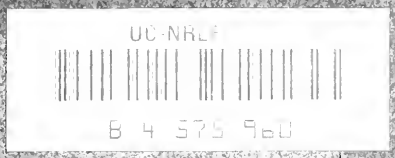
LB
1131
B7



EXCHANGE

SIGILLVM · VNIVERSITATIS · CALIFORNIENSIS

LVX

FIAT

MDCCCLXVIII

EX LIBRIS

EXCHANGE
JUL 27 1917

Variable Factors in the Binet Tests

A DISSERTATION

PRESENTED TO THE

FACULTY OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

BY

CARL C. BRIGHAM

LIBRARY
OF THE
UNIVERSITY
OF
CALIFORNIA

Princeton
Princeton University Press
1917

Variable Factors in the Binet Tests

A DISSERTATION

PRESENTED TO THE

FACULTY OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

BY

CARL C. BRIGHAM

Princeton
Princeton University Press
1917

VARIABLE FACTORS IN THE BINET TESTS

Table of Contents

I.	Introduction	1
II.	Subjects and Methods	8
III.	The Personal Equation	18
VI.	Grade Correlations	37
V.	Sex Differences	65
VI.	Summary	91

I. INTRODUCTION

During the past decade, the Binet-Simon measuring scale for intelligence has received considerable attention, and a large amount of literature has appeared on the subject. No attempt has been made in the following pages to review all the literature on this scale or other systems of intelligence testing. Kite (38) gives an excellent account of the history and nature of the scale. Kohs (41) has assembled a very complete bibliography on the subject up to June 1914. Schmitt (57) gives an historical account of the development of the various attempts to correlate psychological findings with general intelligence, particularly in this country and England. Bobertag (10) and Schmitt both give detailed descriptions and analyses of the individual tests. Stern (62) has devoted a monograph to the collection, exposition and critical analysis of the large amount of data bearing on the problem of intelligence testing, and in another work (61) has assembled the literature of cognate fields. The literature bearing on the Binet scale up to 1912 is largely descriptive of the scale itself, the standard methods of procedure, etc. The more recent literature has been critical and reveals a tendency at the present time for investigators to depart from the methods of the extensive application of the scale as a whole to the more intensive study of the individual tests.

All systems of intelligence tests may be classified as qualitative or quantitative. The qualitative system consists of an aggregation of tests designed to detect the capacities or incapacities of the subject in order to afford the experimenter an opportunity to make a diagnosis concerning the subject's mentality. This method throws the responsibility for the final diagnosis on the experimenter. The system of tests proposed by Healy and Fernald (34) are of this type. Quantitative systems of tests necessitate a final score of some sort, whether that score be in the form of a mental age, a mental quotient, a certain number of points,

a coefficient of intellectual ability, a percentile rank or what not. The essential characteristics of the quantitative systems are the interpretation of the total scores in terms of the age of the subject, and the placing of the responsibility for the final diagnosis on the tests rather than the experimenter.

Binet and Simon's 1905 scale (5 and 6) was of the qualitative type. A series of 30 tests of approximately increasing difficulty was published with directions for their application. The authors reported in a general way that from their experience in examining a few selected normal children of different ages, and other subnormal children in the schools and at the Salpetrière, approximate levels of performance could be found characteristic of the development of normal children of 3, 7, 9 and 11 years chronologically, the performance of idiots, imbeciles and morons corresponding roughly with that of normal children of 3, 7 and 9. Although the reference to chronological ages introduced the quantitative element, at no place were the authors insistent on this point, merely stating that they had found the series of tests exceedingly valuable in diagnosing and classifying defectives, and in their opinion others would also find it valuable.

The 1908 scale (7) was quantitative in character owing to the introduction of the concept of "mental age". It included a list of 56 tests grouped according to ages from 3 to 13, each group containing from four to eight tests. Most of the tests of the 1905 series were included, the additions including in a large measure tests of a scholastic nature. The authors gave directions for applying the series and for computing the resultant "mental age". A child testing three years below his chronological age was to be considered defective.

Although the scheme of the 1908 series was entirely quantitative, the authors did not discard the qualitative idea, and they cautioned against the application of the scale in the manner of a measure of height or weight. The border line between the idiot and the imbecile was fixed by the ability to use and comprehend spoken language. The imbecile was differentiated from the moron by the use of written language, illiteracy being di-

ferentiated from imbecility by certain tests. The authors stated that the moron could be defined only in terms of the environment in which he lived, and they considered six tests important in differentiating the moron from the normal individual of the Paris population. Any system of tests which throws more weight on some tests than on others in making a differential diagnosis is fundamentally qualitative in kind, for the responsibility is placed not on the score but on the judgment of the experimenter. The idea of a quantitative measuring scale of intelligence however met with instant favor. The interest that actuated the psychologists of the "early nineties" to correlate the measurements of reaction time, motor ability, sensory discrimination, etc. with intelligence was revived. The scale was translated into several languages and applied to individuals of many classes and types.

In 1911, the authors published a revised scale (8) in which many of the tests of scholastic ability were discarded, and the remaining tests shifted about so that there were five tests for every year except one from III to X with similar groups for "twelve year", "fifteen year" and "adult" mentality. In the same year, Binet published an article (4), his last word on the subject, in which he discussed many of the criticisms which the scale had received, and again sounded the note of warning against the mechanical interpretation of results. However, as one traces Binet's thought on the subject through his writings, he may see the idea of a qualitative system of tests gradually dropping into the background, and more and more weight placed on the "scientific" (quantitative) measure of intelligence.

That Binet did not depart entirely from the qualitative standpoint is shown by his discussion of the test of comprehending difficult questions. "Sometimes after an examination one hesitates on a diagnosis. The child has failed in one or two tests, but this does not seem to be convincing. Failure to give the day and date and the months of the year are excusable errors, which may be caused by distraction or by lack of education. But the questions for comprehension dissipate all doubts. We recall

several instances when teachers brought us children, desiring to know whether or not they were abnormal; occasionally, in this way they set a trap for us, but we did not object, it was fair play. Our questions for comprehension decided us every time. We remember one child who was very slow in answering as though dull, his face was expressionless and unprepossessing; he knew neither the day nor the date, nor what day comes after Sunday, and he was 10½ years old; his reading was syllabic. But when we asked question 5: Why do we judge a person by his acts rather than by his words? he gave the following answer: Because words are not very sure and acts are more sure. This was enough—our opinion was formed, that child was not so bad as he seemed." (Town's (72) translation, page 48.)

The popular interest that was manifest before the advent of the 1911 scale was tremendously reinforced in this country by Goddard's (30) publication of the results of the application of the scale to "two thousand" non-selected school children in Vineland, N. J. Popular interest increased rapidly, and the scale continued to have wider and wider application in the hands of less and less experienced investigators. The concept of "mental age" was exceedingly easy of comprehension, no apparatus was needed, and the scale has now become the common property of all. This development or overdevelopment has taken place in spite of the warnings of the authors themselves and the psychological fraternity in general. The very fact of overdevelopment however is striking evidence that persons interested in the social sciences need a quantitative scale for measuring intelligence.

The question whether the Binet scale is an accurate measure of intelligence can be decided only by the study of the individual tests and the factors underlying them. A study of this sort will show the errors that underlie the total score or "mental age", and at the same time will show the direction in which the correction of the scale should take place. The proper understanding of the individual tests involves the theory on which the measuring scale was constructed.

The method which Binet and Simon used in constructing their

measuring scale of intelligence was entirely empirical. A large number of tests were given to children of a certain social status. Certain tests could be shown to be correlated with age, and in the authors' opinion were correlated with intelligence. The fact that at a certain age a test could be passed by a certain proportion of the subjects was taken to mean that the test in question was characteristic of that age. Tests that were characteristic of the same age level were then combined into one age group. In this way a scale was built up with a number of tests for each age group. By a certain arbitrary system of scoring the reactions of a subject to all or part of the scale of tests, the "mental age" of the subject was obtained. The comparison of the "mental age" with the chronological age of the subject would show him to be advanced, at age or retarded, and the amount of acceleration or retardation would afford a quantitative index of his intelligence.

A person could construct a scale on the same basis and arrive at an age score using entirely different tests. A scale could be constructed containing tests of height, weight, vital capacity, strength of grip, circumference of the head, etc. and the results interpreted in terms of age. In this case however the age obtained would be more physical than mental. A scale of tests could also be constructed which involved the subject's knowledge of geography, spelling, history, grammar, etc. but in this case the resulting age would be determined very largely by the amount of training the subject had received.

The assumptions that a child at a certain age should weigh 25 pounds, at another age 50 pounds, etc., that a child can repeat 3 digits at one age, 5 digits at another and 7 digits at another, and that a certain percentage of children at one age can enumerate the months, and a higher percentage at another age, differ only in the possible determiners to which the growth may be referred. In the first case the growth is referred to certain physiological processes which are supposedly independent of intelligence and training. Binet believed that the principal determiner of growth in the last two cases was intelligence, but the possibility

remains that they might be more or less independent of intelligence, and more or less dependent on training and other variable factors.

The principle on which the scale was constructed involves three assumptions, (1) that the individual tests are correlated with age, (2) that the individual tests are correlated with intelligence, and (3) that intelligence is correlated with age—three distinct assumptions any one of which does not necessarily involve the others. The purpose of this investigation is to study the correlation of the individual tests with age, to determine the variable factors that might operate on the tests to produce an apparent correlation with age that was not a real correlation, or that might alter the real correlation in some way.

There is a possibility that an error might occur in the statistical treatment of the results, so that figures which would apparently indicate a correlation with age of a certain degree might actually represent a correlation of another degree. Another variable factor is the personal equation of the experimenter, who might alter the procedure in giving a certain test so that the correlation of that test with age might be different from the correlation obtained by another experimenter. If the subjects of various ages had received different school training, this difference might introduce another factor which would vary independently of the age of the subjects. If the tests used depended on any inherited or acquired differences between the sexes, then the correlation of the tests with age might be different for the two sexes. If any or all of the variable factors mentioned prove to be present in the correlation of the tests with age, then certain allowances will have to be made for these factors in making a diagnosis of the subject's intellectual ability on the basis of his total score or "mental age", and the scale becomes qualitative rather than quantitative.

At the Fourth International Conference for School Hygiene held in Buffalo in the summer of 1913, several persons of unquestioned authority in the field of mental tests held an informal

conference on the Binet-Simon scale, reporting the results in 1914 in the form of recommendations and suggestions (15). The question, "How much is the outcome of the testing influenced by the personal equation, both of the examiner and examinee?" was answered, "Undoubtedly there is some influence and it may be a serious source of error." Another question, "How much do previous environment and school training effect the outcome of the tests?" was left unanswered by the opinion, "The experimental evidence thus far available is conflicting. Further investigation is needed." The question, "Should the scale be divided, in the upper years at least, to furnish separate standards or separate tests for the two sexes?" was answered, "We do not know, and recommend this a subject for investigation." The following study is in part an attempt to answer these questions.

The method used in this study is that of studying the individual tests, disregarding entirely the total score or "mental age". There are at present so many revisions and editions of the Binet scale, that the term "mental age" has no meaning outside of the particular scale in question. The tests that are used in the various standardizations are however approximately the same, so that conclusions concerning the factors underlying the individual tests have a wider significance than those drawn from the "mental ages". Furthermore variable factors in the individual tests may balance each other in the total score so that their influence might be obscured.

The subjects and methods will be described first, and in connection with the methods of treating the results a statistical error will be pointed out. The problems of the personal equation, grade correlations and sex differences will then be taken up in detail.

II. SUBJECTS AND METHODS

Subjects

The data which are here analysed to determine the influence of the personal equation, of grade training and of sex differences, are derived from all the boys and girls below the seventh grade in the Princeton, N. J., Model School. This group includes 422 subjects of the following age distribution,—

Chronological Ages.

4	5	6	7	8	9	10	11	12	13	14	15	16
4	17	62	52	56	42	53	49	36	32	11	6	2

Each of the first six school grades was divided into a plus and minus grade, the latter division being under a different teacher, and containing those who were either backward, or, on account of illness, change of school, or for reasons not necessarily related to their mental development, were not sufficiently advanced to perform the work of their grade. The school also contained a special class for defective and exceptionally backward children. The subjects were distributed in the school grades as follows,—

School Grades.

Spec. Kind.	I—	I+	II—	II+	III—	III+	IV—	IV+	V—	V+	VI—	VI+
18	32	38	51	12	40	12	45	15	35	15	49	11	49

39 or 9.2% of the subjects were children of non-English speaking parents, this group including 6.6% of the children in the Kindergarten and first six regular grades, and 15.7% of those in the special class and minus grades.

The selection of subjects is only fairly typical of the general run, for Princeton has no manufactories. The children examined came, for the most part, from the homes of laborers, domestics, artisans, farmers, tradesmen, clergymen and college professors. The selection is atypical in that none of the children came from homes of the manufacturing class, while an unusually large pro-

portion came from the homes of those engaged in domestic, personal, and professional service.

Tests

The scale used was Goddard's (28) 1911 revision of the Binet-Simon scale. The methods used in giving the tests were, as far as possible, the same as those outlined by Goddard in the original revision, incorporating the rules and suggestions for standardized scoring published by that writer (29) in 1913. The methods used will not be discussed in detail, for the data are not used in obtaining age norms and standards for children generally. For the analysis of the data in terms of grade and sex it is not necessary that the procedure should be absolutely standardized, but that the experimenters who gave the tests should have used the same procedure. Differences in the technique of the experimenters will be discussed in the chapter on the personal equation.

One variation from the usual procedure was adopted. In no case did the experimenter know the chronological age of the child being tested. The influence of any prejudice or bias on the part of the experimenter is therefore eliminated from the problem of the correlation of the tests with age. The three experimenters who gathered the material in the spring of 1913 examined the sixth grade first and the remaining grades in decreasing order. During the school year 1913-1914, the fourth experimenter examined all children at that time in the kindergarten and first grades, and others who were not examined in the spring of 1913.

The tests in the "three year", "four year", "five year", "fifteen year" and "adult" groups were given so infrequently that the data from them are not treated. The tests used are as follows. The figure at the right shows the total number of times each test was given.

AGE VI

1. Distinguishing between morning and afternoon	108
2. Defining in terms of use	333
3. Executing three commissions	100

4. Showing right hand and left ear 107
5. Choosing the prettier of given faces 117

AGE VII

1. Counting 13 pennies 217
2. Describing pictures 219
3. Indicating omissions in pictures 217
4. Copying the diamond (in pencil) 225
5. Naming four colors 218

AGE VIII

1. Comparing remembered objects (butterfly and fly) 271
2. Counting backwards from 20 to 0 251
3. Enumerating the days of the week 277
4. Counting stamps 258
5. Repeating 5 digits 413

AGE IX

1. Making change 271
2. Defining in terms superior to use 333
3. Giving the day and date 307
4. Enumerating the months 284
5. Arranging five weights 334

AGE X

1. Recognizing pieces of money 282
2. Copying designs from memory 252
3. Repeating 6 digits 413
4. Comprehending easy and difficult questions 250
5. Using three words in sentence (two ideas) 279

AGE XI

1. Detecting absurdities in statements 226
2. Using three words in sentence (one idea) 279
3. Giving 60 words in three minutes 233
4. Giving rhymes with day, mill and spring 213
5. Reconstructing dissected sentences 190

AGE XII

1. Repeating 7 digits 413
2. Defining abstract terms 144
3. Repeating a sentence of 28 syllables 169
4. Resisting suggestion (length of lines) 203
5. Solving problems from various facts 123

The tests in the "six year" group, with the exception of defining in terms of use, and the tests in the "twelve year" group, with the exception of repeating 7 digits, were given so infrequently or so irregularly that the data from them could not be treated. The apparatus used in the test of arranging five weights was not constant throughout the experiment, the standard cubes

and weighted pill boxes being used at different times by different experimenters. On this account, the data from this test are not included in the subsequent discussion.

Methods of Treating Results

The chronological age of each subject was taken as that at the last birthday, one tenth of a year being allowed for each 36 days beyond the birthday. The subject that was 10 years and 35 days would be rated 10.0 years, while ten years and 36 days would be 10.1 years. A subject one day short of 11 would be rated 10.9 etc. The teachers of each grade submitted the dates of birth of all pupils after the grade had been tested. These data were later checked up from the entrance cards. Since the purpose of this study is to analyze the factors involved in the individual tests, no "mental ages" or total scores were figured. The classifications of the subjects are all made independently of the tests.

Two measures of central tendency will be used in the subsequent discussion, the average and the median. The measure of variability from the average, that will be used, is the mean variation (or average deviation), the average of the differences, regardless of signs, between the separate measures in the series and the average of the whole series. The measure of variability from the median that will be used is the semi-interquartile range (Q), or half the difference between the measure with three times as many measures above as below it and the measure with one third as many measures above as below it, i. e. half the difference between the 25 percentile, and the 75 percentile. Any coefficients of correlation used will be stated in terms of the formula applied. The reader is referred to Thorndike (70) for the discussion and explanation of the statistical measures used.

The measures of ability in most of the tests are in the "all or none" form—the tests are either passed or failed. The only measure that can be obtained from data of this sort is the percentage that an ability is present in a defined group. This method of treating the results has as many "pit-falls" as the tests themselves. Before undertaking the analysis of the Prince-

ton data to determine the effect of the personal equation of the experimenter, and the age, grade, and sex of the subject upon the results of the individual tests, it is necessary to consider an error which underlies incomplete data, or those data derived from experimenting in which every test is not given to every subject.

No uniform instructions were given to the experimenters concerning the order in which the tests should be given, nor the number of tests that should be tried. The experimenters attempted to determine the mental age of the child according to the scale. In doing this they would start with some test which they considered would be interesting to the child, and, at the same time, well within his reach. The tests given first were usually those of describing pictures and arranging five weights. The experimenter would then gradually explore the subject's range of ability, varying the order of the tests so as to maintain the subject's interest, and to ward off fatigue. In this way the experimenter would eventually establish the basal age of the subject (that age in which he passed all five of the tests), and by the end of the examination would have tried all the tests above the basal age which, in his judgment, there was any possibility of the subject's passing. This method of experimenting will be called incomplete. The other method of experimenting, in which a certain number of tests are adopted and all of the tests are tried on each subject, will be called complete. Each experimenter in the Princeton investigation averaged 19 or 20 tests to a subject. In the Trenton investigation all the tests were given to all the subjects.

The incomplete method is more desirable from the standpoint of the subject who is not unnecessarily fatigued, and from the standpoint of the experimenter, as well, who saves in the expenditure of time and energy. However, the data derived from the incomplete method are subject to an error, which, unless it is properly considered, will completely vitiate the results.

When the experimenter does not try a test above the basal age because he believes that the subject will not pass it, he implies that the subject will fail it. This amounts to a failure,

for the subject receives no credit. However, a failure of this sort, due to the experimenter's assumption, is not the same as an actual failure in which the test is tried, for there is always the possibility that the assumption was unjustified. In like manner when the experimenter does not try tests below the basal age, he actually gives credit for passing the test without the actual trial.

In some cases the assumption on the part of the experimenter is quite justified. Obviously if a subject can make change, he can count up to thirteen; if he can repeat seven digits, he can repeat five and six digits; if he knows the names of the months, he will know the days of the week; and, conversely, if he cannot repeat the days of the week, he cannot repeat the months. Other assumptions are less justifiable. Since very intelligent persons, lacking in particular sorts of abilities, might fail in tests such as drawing the design from memory or arranging five weights, there is no reason for supposing that a subject making basal "eleven" or "twelve" will pass these tests. At the same time there is no reason for assuming that a subject failing to establish basal "seven" for instance, will fail to pass a test such as the line suggestion test in "twelve". The assumptions of the experimenters, then, are more or less justifiable and it is impossible to estimate the amount of the justification, since this is dependent on the nature of the individual tests.

The manner in which this error works out in the statistical treatment of the results may be shown by examining any test which has been tried through a number of chronological ages. Table 1 shows the results of the 60 word test obtained from subjects 7 to 13 years of age.

TABLE NO. 1

Analysis of the Results from the Test in Naming 60 Words in 3 Minutes.

Chronological ages	7	8	9	10	11	12	13
No. of times given	11	18	25	42	44	31	28
No. of time passed	4	10	10	24	34	19	21
Actual percentage passed	36%	56%	40%	57%	77%	61%	75%
Total number of subjects	60	52	42	54	48	36	28
Percentage of subjects to whom test was given	18%	35%	60%	78%	92%	86%	100%
Theoretical percentage passed	7%	19%	24%	44%	71%	53%	75%

An example will make the above table clear. The 60 word test was given to 11 subjects, age seven, 4 of whom passed. In all there were 60 subjects at this age, so that the 11 subjects to whom the test was given constitute but 18% (and probably the brightest 18%) of this whole number. The percentage passed would have been 7% had the test been given to all 60 subjects, and had all the subjects failed who the experimenters assumed would fail if they gave the test. The true per cent. which represents the ability of non-selected seven year boys and girls in passing the 60 word test therefore lies somewhere between 7% and 36%, probably nearer 7%. An accurate estimate of the real per cent. which will represent this ability is, however, impossible. In like manner, the ability of the 8 year subjects is represented by a percentage somewhere between 19% and 56%.

As the proportion between the number of subjects in the group and the number actually tested increases, the disparity between the actual and theoretical percentage passed becomes less, or, in other words, the results which express the ability of a group become more reliable as the number of individuals actually tested as a sample of this group becomes larger. The higher the percentage given, the more reliable the percentage passed, when the reliability is measured by the difference between the actual percentage passed and the theoretical percentage passed.

The source of error mentioned causes great difficulty in comparing the results of different investigators. For example, it is desired to compare the results of Terman and Childs (66) and Dougherty (23) with those of this investigation on the 60 word test. Table 2, derived from their published results, shows the percentage that the test was given of the number of times it was possible to be given, (%G), the actual percentage passed, (A%P), and the theoretical percentage passed, (T%P), or that percentage passed that would have resulted had all of the subjects failed, who it is necessary to suppose would have failed, had the test been given all the possible number of times.

TABLE NO. 2

Analysis of the Results of Three Investigators on the 60 Word Test.

Age	This investigation			Terman and Childs			Dougherty		
	%G	A%P	T%P	%G	A%P	T%P	%G	A%P	T%P
7	18	36	7	14	50	7			
8	35	56	19	47	35	16	15	0	0
9	60	40	24	86	57	49	35	60	21
10	78	57	44	100	67		78	53	41
11	92	77	71	98	83	82	89	79	70
12	86	61	53	97	82	80	91	95	87
13	100	75		100	94		94	88	83

It is very difficult, if not impossible, to make a comparison of these results shown in Table 2 for the years 7, 8 and 9. The ability of Terman's 7 year group is represented by a figure somewhere between 7% and 50%, while that of the 8 year group falls somewhere between 16% and 35%. Dougherty's 9 year group falls between 21% and 60%. In the older years where the results have greater reliability, it is probable that the discrepancies between the investigators could be accounted for on the basis of the inferiority of the selection of the older subjects in this investigation, the other investigations including children from the seventh and eighth grades.

In order to make a comparison between investigators, it is necessary to express the results in terms of a percentage or a proportion. The expression of the ability of a group by a percentage or a proportion is inaccurate if the data are incomplete, and in order to judge the accuracy of the data, it is necessary to know the degree of completeness. Unfortunately, the results of most of the investigations on the individual tests are not published in a form that enables one to estimate the accuracy of the data. The writers who have published their data in a form that will admit of this treatment, have not treated the sexes separately. On this account, the writer will not attempt a systematic comparison of the results of this investigation with those of other experimenters.

Before analysing the Princeton data the following problem should be answered:—What proportion of a given group must actually be tested for an ability in order that the results may be

considered as typical of the ability of the whole group? The proper proportion to select as typical of any one group depends upon the characteristics of the group itself. If the members of a group are similar, a smaller proportion would stand for the ability of the group than would be necessary for a group composed of unlike individuals. A smaller number of individuals would be necessary to stand for the ability of all the 12 year boys in the sixth grade, for example, than for all the 12 year boys coming from a great many grades. This proposition operates directly counter to actual practice, for the members of a group of similar individuals will be given similar tests, while unlike individuals will receive different tests, inasmuch as the experimenter adapts his procedure to the need of the individual being examined. The proposition actually means, then, that selected results from incomplete testing are more reliable than non-selected results, if each group has the same range of testing. The proportion of a group that must be tested to stand for the whole group will also vary from test to test. In some tests of particular abilities, no proportion will accurately stand for the whole group—the entire group must be tested. In other tests that are easy for the group, the results of a very small proportion would not be altered by examining the remainder of the group.

The problem of deciding what proportion of a given group must actually be tested for an ability in order that their results may be considered as typical of the ability of the whole group has, therefore, no answer in the work. The writer will decide arbitrarily what the proportion will be. The actual magnitude of the proportion between the number actually tested and the number in the whole group (the percentage given) will always be published as an index of the reliability of the percentage that the group passes the test in question.

It is not possible to obtain reliable results showing the growth of an ability with age, if the data on which the results are based are of the incomplete sort. A test for any age will be given to a superior selection of subjects below that age, and an inferior selection of subjects above that age, so that the growth curve

will appear flatter than it actually is. For this reason, the Princeton data may not be used for the purpose of standardizing age norms.

Binet (4) recognized the fallacy of calculating proportions from the actual number of times a test was given and passed when the test had not been given all the possible number of times. In calculating the proportions from Levistre and Morle's data, Binet used what the present writer would call the "theoretical proportion passed".

It has been shown that the reliability of the theoretical percentage passed rests on the accuracy of the experimenters' assumptions, and that according to the nature of the tests and the character of the groups to which they are given these assumptions vary from complete certainty to absolute uncertainty. Inasmuch as these assumptions are not equally certain, the conclusions drawn from them are not equally certain, and the logic of scientific method demands that an investigator establish the degree of certainty of his conclusions. In this case the measure of the degree of certainty is the magnitude of the percentage given.

The use of the theoretical percentage passed without reference to the percentage given ignores the dictum that an investigator establish the degree of certainty of his conclusions, and sets up all conclusions as equally valid, a procedure which in actual practice results in making all conclusions equally invalid when the fact of degrees of certainty is admitted. The investigator who draws conclusions from incomplete data should always state the percentage given and the actual percentage passed. This much at least is experiment. The only legitimate use of the theoretical percentage passed is when it is compared with the actual percentage passed as a probable limiting value. The theoretical percentage passed alone has no claim to reliability.

III. THE PERSONAL EQUATION

Before attempting to correlate the individual tests with age, grade and sex, it is necessary to demonstrate the presence or absence of the effect of the personal equation. By the term "personal equation" is meant the complex of variable factors which are independent of the mental make-up of the subject and the environmental conditions at the time of the examination. The term includes such widely different factors as the experimenter's ability to obtain the cooperation of the subject, his procedure in giving the tests, his criteria in deciding whether a subject's response should pass or fail, and the tests used, insofar as the selection of tests and the construction of the apparatus were occasionally left to his discretion, apart from the uniform procedure.

The only method of detecting the influence of the personal equation in most of the tests is that in which the responses of similar groups of subjects to different experimenters are compared. On account of the wide variations in the character of the subjects examined, it is not possible to compare similar groups. On some tests, however, it is possible to determine the effect of the personal equation independently of the method of group comparison. The results of the tests that may be studied independently will be discussed at some length, in order to demonstrate the fact that certain tests are susceptible to this influence.

The examinations of the Princeton subjects were made by four experimenters, called for convenience A, B, C and D. None of the experimenters was highly trained in giving the tests, although they had all been trained in the methods of psychological experimentation, one experimenter being an assistant professor of psychology, and the other three graduate students of psychology of at least one year's standing. B, C and D performed their experiments at the same time, in the spring of 1913, while A experimented one year later. B, C and D studied

the scale together so that it was possible to secure a correspondence in method. At the close of practically every day's testing, B, C and D would confer on the questions brought out by the day's work, and as far as possible would adopt uniform methods of procedure and scoring. A was subsequently trained in these same methods.

In spite of the attempt to adopt uniform methods, there were a few tests which always caused difficulty, and concerning which the experimenters could reach no definite agreement. One of the tests that caused the greatest difficulty was that of defining in terms of use and in terms superior to use. The hierarchy of responses to this test could be fairly arranged as follows. To the question "What is a chair?" the following typical responses would be obtained,—1, "A chair is a chair." 2, "This is a chair." 3, "A chair is to sit on." 4, "A chair is what you sit on." 5, "A chair is a thing you sit on." 6, "A chair is a piece of furniture you sit on." 7, "A chair has four legs, a back, etc." 8, "A chair is a piece of furniture with four legs, a back, etc." Any of the objects for which a definition is asked (fork, table, chair, horse, mother) may be defined by repetition, by demonstration, by indicating the use to which it is put, by showing the class to which it belongs, by describing its parts, or by the combination of any or all of these methods.

The only problem is to decide, arbitrarily, how definitely the class must be indicated (i. e. by "what", "a thing" or "a piece of furniture") in order to have the definition considered as one of classification. The rule adopted in this study was to consider "thing" as indicating the class. Nos. 1 and 2, definitions by repetition and demonstration, received no credit in "six years". Nos. 3 and 4 were given credit in "six years" as definitions by use, and nos. 5, 6, 7, and 8 were given credit in "nine years" as definitions in terms superior to use.

In studying the ranks given to the responses of the subjects in this test, it was found that the experimenters did not record the responses all of the time. A gave the test 94 times, and recorded the responses 66% of this number. B gave the test 98 times, recording the subject's answer 67% of the time. C

gave the test 65 times, and recorded the answer in 95% of the cases, while D gave the test 76 times and recorded the response only once.

By ranking the recorded responses of A, B and C according to the rules shown above, it is possible to obtain an estimate of the relative severity of their criteria in marking these responses plus or minus. 19% of A's definitions were corrected, the correction in all cases being from minus to plus. 11% of B's definitions were corrected, all of the corrections being from plus to minus. 17% of C's definitions were corrected, three fourths of them being changed from plus to minus, and one fourth from minus to plus. C's standards changed during the course of the experiment, so that at first, with older subjects, he was too lenient, while later, with younger subjects he was slightly too severe. The tendencies of A and B remain constant throughout the experiment, A marking too severely and B slightly too leniently. The differences between the experimenters hold constant for both sexes. The experimenters agreed on all definitions by use, the cases of disagreement coming on the definitions superior to use.

One test in which variations between the experimenters might be expected is that of copying the diamond. In this test, although the apparatus and procedure were the same, the experimenters had very little to guide them in forming their judgments of passed and failed. The instructions given ("The result is considered satisfactory if it would be recognized as intended for a diamond shaped figure"), and the examples published furnish very vague criteria.

In order to determine the effect of the personal equation of the experimenters in giving credit on this test, all of the reproductions of the diamond obtained in the Princeton and Trenton experimenting, (311 in number), were first transcribed and then ranked. On the sheet containing the copy only the subject's number was placed, so that the person ranking the reproductions was in ignorance of the experimenter by whom it was obtained, the mark that the experimenter had given it, and the age, grade, sex, etc. of the subject. The 311 diamonds were then classified

into six groups by one observer. The classification, at best, was vague and indefinite, but it represented the unbiased judgment of a single person. Inasmuch as the reproductions were classified and re-classified a great many times, small errors in the classification would be counterbalanced.

The first group contained fairly accurate reproductions of the original, diamonds of approximately the same size as the copy, having the sides and opposite angles nearly equal, and with a proper proportion between length and width. The second group contained figures inferior to those of the first group in size or symmetry, but representing a fairly high grade of ability. The reproductions that were less symmetrical than those of the second group were classified in the third and fourth groups. Figures showing some inequality between length and width were classified in the third group, while those of approximately unit proportion, square shaped figures, were classified in the fourth group. The reproductions placed in the fifth group were figures less symmetrical than those of the fourth group, and figures which had curved sides and rounded corners. The sixth group contained all figures which it would have been difficult to have recognized as intended for a diamond, figures having three, five or more sides, circles, elipses, unfinished lines and eccentric figures.

The above classification did not offer an opportunity for a sharp grading between one group and another, but in general, the reproductions placed in the various groups from the first to the sixth, represented a decrease in the ability to copy the diamond. The justification of the method was not in the accuracy of the classification, but in the fact that the material was all classified by one observer (B), in such a way that he was in ignorance of the original rank that had been given the reproduction, of the experimenter who graded it, and of the character of the subject.

16% of the reproductions were classified in the first group, 21% in the second group, 20% in the third group, 17% in the fourth group, 9% in the fifth group, and 17% in the sixth group. (The irregularity of the distribution is due to the presence of the diamonds drawn by the Trenton subnormal group.)

After classifying all of the reproductions the ranks given to them by the different experimenters were then compared with the group in which they were classified. That the sliding scale classification used represented real differences between the reproductions is shown by the relative certainty of the experimenters' judgments. None of the reproductions classified in the first and second groups were ranked as failed by the four experimenters, while only one reproduction in the third group was ranked minus. 18% of the fourth group, 45% of the fifth group and 77% of the sixth group were ranked as failures. All of the sixth group diamonds that were ranked plus (23%), were so ranked by one experimenter, A.

To obtain a general estimate of the relative severity of the experimenters' criteria in making their judgments of passed or failed, the diamonds obtained by each experimenter from boys and girls were classified according to rank, plus or minus, and according to their group in the classification. From this it was possible to obtain an estimate of the passing mark of each experimenter. For example, the boys of experimenter B passed the test 72% of the time according to his ranking. Had B given credit for the first five groups and failed only the reproductions in the sixth, i. e., had his passing mark been the fifth group, 88% would have passed. Had his passing mark been the fourth group, 81% would have passed. If it had been the third group, 72% would have passed, while only 56% would have passed had it been the second group. Since 72% of B's subjects actually passed the test, his passing mark was the third group—in the long run, he would pass all diamonds in the first three groups and fail all in the last three. The differences between the experimenters on this basis are quite marked. The passing mark for C and D was the fourth group, while A's passing mark was the fifth group. B was the most severe, and A was the most lenient, with C and D between the two. The results were the same for both sexes.

Another test in which the influence of the personal equation might be looked for is that of copying designs from memory. The experimenter must here use his own judgment in marking

the designs passed or failed. Very little guidance is given by Binet's rule, which reads, "The test is considered passed when one of the designs is reproduced exactly, and half of the other is correctly drawn", or by the interpretation of this "half right" as applying "when two component parts are transposed or one component part omitted".

In order to test the experimenters' judgments in ranking this test, a scoring system was devised, which may be explained by reference to Figure 1, which gives the original copy and various duplicated portions. In scoring the reproductions of the pyramid section, 5 points were given when the reproduction of the asymmetry of the figures was nearly exact, as in no. 1, 4 points for a less perfect reproduction as in no. 2, and 3 points for a reproduction in which the rectangle fell in the center of the figure, as in no. 3. 1 point was deducted from this score for each failure to connect the corners of the rectangles as in no. 4 (which is modified from no. 3 and would therefore receive only 1 point), and no credit was allowed for "boxes" (no. 5), and other eccentric figures.

In scoring the more complicated design, 4 points were allowed for each of the "posts", ABCDE and JKLMN, or no. 6. 2 points were deducted for turning them in the wrong direction as in no. 7, (which is "post" ABCDE turned in the wrong direction), 2 points for failure to make the line AB penetrate DE as in no. 8, so that a combination of these errors, as in no. 9, would receive no credit, along with other eccentric reproductions as in nos. 10 A, B, C, D, E and F. 1 point was given for each of the lines EF and IJ, and 5 points for the "hump", FGHI. A continuous line from E to J as in no. 11 would therefore receive no credit, while a division of the lines, without the portion FGHI, as in nos. 12 or 13, would receive 2 points. An accurate reproduction of the portion EFGHIJ, as in no. 14, would receive full credit for all parts, 7 points, no credit being allowed for eccentric reproductions of the "hump" as in nos. 15 A, B, C and D.

The maximum credit for the test is 20 points, divided between the two figures on the proportion of 5 to 15, a fair proportion

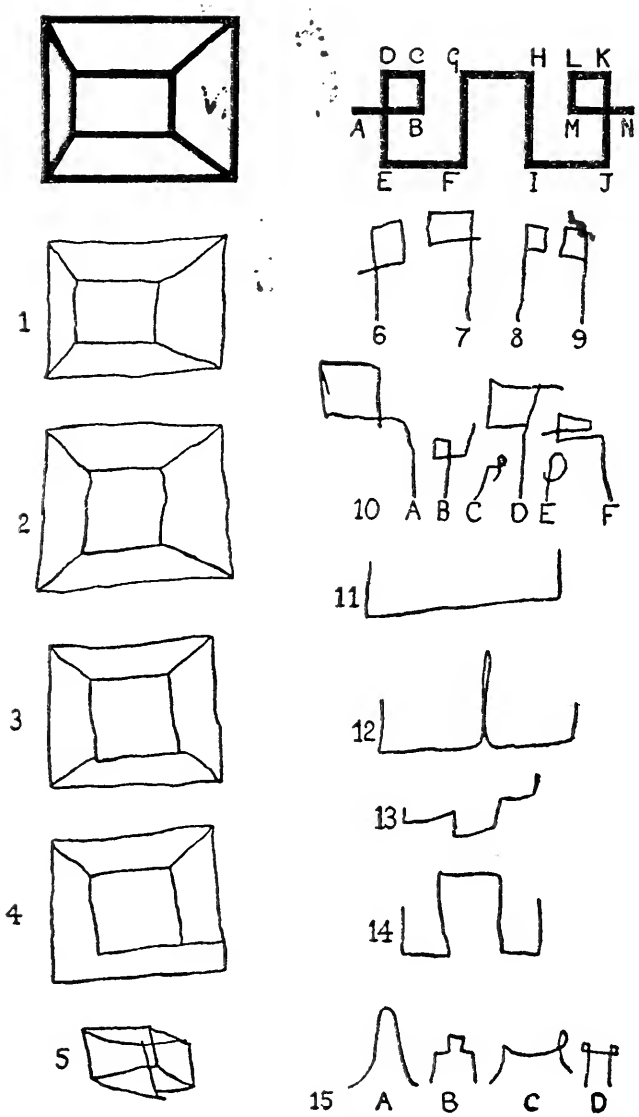


FIG. 1. *Method of Scoring Test of Copying Designs from Memory*

(in the writer's opinion) according to the relative difficulty of the parts. A design with "one component part omitted" would be scored 13 points according to this system, and one with "two

component parts transposed", 16 points, provided that the reproductions of the pyramid section were perfect in each case.

All the reproductions of the designs obtained from the Princeton and Trenton experimenting were then scored according to this system. The score of each subject of each experimenter in the Princeton series was then compared with the experimenter's ranking, which was recorded on the same sheet, and which was not seen at the time the designs were graded by the point system. From the number of times the test was given, and the number of times it was marked passed by the experimenter, the percentage passed was obtained for each experimenter for both sexes. The scores from all the designs from 0 to 20 were then classified according to the judgment passed or failed as given by each experimenter on subjects of both sexes. It was found that there were certain ranges where the experimenters' judgments coincided accurately, i. e. in the very low scores and in the very high scores. A certain range existed, approximately from 10 to 15 points, in which the same results would sometimes be ranked as passed and failed by the same experimenter at different times.

It was possible, however, to obtain a general estimate of the experimenters' criteria by a method similar to that used in the study of the diamond test. For example, B gave the test to boys 48 times, passing 40% of them. Had his passing mark been 18 (i. e. had he passed all subjects whose designs scored 18 points or better), 21% would have passed. Had his passing mark been 15 points, 35% would have passed. Had it been 13, 42% would have passed etc. B's passing mark would therefore fall between 13 and 15 points. In this way, by calculating the percentage passed at each score for each experimenter for both sexes, it was possible to obtain the passing mark of each group. The passing marks coincided very closely except in one case. With one exception the passing marks were around 12, 13, 14 or 15 points, for the boys and girls of all experimenters, i. e. the experimenters would, in the long run, rank all below this level minus and all above this level plus. The degree of correspondence was quite remarkable considering the fact that the

experimenters had very little on which to base their judgments.

The one exception is both striking and suggestive. C's passing mark for boys was 15 points, for girls 8 points. In order to receive a plus from C, boys would have to draw a much more accurate design than girls, or, in other words, a very faulty reproduction drawn by a girl would receive credit, while the same reproduction if drawn by a boy would invariably be failed. This deviation rests on a small number of cases. A gave the test to 24 boys and 21 girls, B to 48 boys and 33 girls, C to 28 boys and 22 girls, and D to 36 boys and 31 girls. A's results, although resting on a number of cases as small as C's, show no such deviation as those of the latter. On account of the small number of cases, this finding cannot be considered definite. It does, however, suggest the possibility of a difference in the experimenters' reaction to the sexes. An experimenter may show greater leniency to one sex than to the other, so that a supposed sex difference may be the results of an experimenter's reaction to the sex, rather than the sex's reaction to a test.

The test of using three words in a sentence ("Philadelphia, money and river") was given 279 times, and the sentences given by the subjects were recorded over half the time. Experimenter A gave the test 53 times, recording the result 36% of the time. B gave the test 95 times, recording the answer in 92% of the cases. C gave the test 56 times, recording the answer 23% of that number, and D gave the test 75 times, recording the response in 43% of the cases.

To obtain a check on the accuracy of the experimenters' scoring of this test, all of the recorded sentences were transcribed so that they could be studied and ranked without reference to the subject or the experimenter. The 162 recorded sentences were then marked plus or minus by one observer (B). This ranking was checked several times and then compared with the original ranking.

There was no disagreement between the judgments of the four experimenters and the one impartial observer in marking responses for the "ten year" credit. In marking for the "eleven year" credit, there were 8 disagreements out of the 162 judg-

ments, the 8 variations being evenly distributed among the experimenters. It may be concluded, then, that the influence of the personal equation is absent in this test, although there is ample opportunity for variation.

The detailed study of the foregoing tests has shown that the personal equation of the experimenters has a marked effect on the results of some of the tests. In the subsequent correlation of the tests with grade and sex the corrected score of these tests will be used. Only those definitions will be used which were recorded by the experimenters, and the ranking of the one observer will be followed. All reproductions of the diamond in the fifth and sixth group will be scored as failed, the others as passed. A reproduction of the designs scoring 15 or more points will be ranked as passed. The corrected results of the sentence test will be used.

To show that the effect of the personal equation of the experimenter is present or absent in the tests on which there is no actual record of the subject's response, is a more difficult problem. The most reliable method of showing the influence of this factor is that in which the reactions of similar groups of subjects, examined by different experimenters, are studied. The greater the similarity of the groups the more reliable the results. If two experimenters each examined 50 boys of 12 years of age from the sixth grade, their results should compare closely, and any difference could immediately be referred to a difference in the personal equation. However, if one examined boys from this grade and the other girls, the variations might be explained on the basis of sex differences. In the same way the results may vary with the age of the subject, and with his grade and nationality.

It is not possible in this study to obtain groups of a sufficient degree of similarity, in spite of the small number of children of non-English speaking parents, and the fact that the sexes may be treated separately. The subjects vary in age from 4 to 16, and in grade from the kindergarten to the sixth grade. A examined a very much younger run of subjects than B, C and D. The data of the four experimenters were treated by three meth-

ods, by comparing the per cent. that all boys and girls of each experimenter passed each test, by comparing the per cent. that selected subjects of each experimenter passed each test, and by comparing the per cent. that all subjects from 5 to 9 and from 10 to 13 passed each test. The sexes were separately treated in each method. None of the methods proved satisfactory, and it was found to be impossible to obtain an accurate quantitative estimate of the effect of the personal equation on each test. In certain of the tests, however, there were known differences of procedure which might have influenced the results, while the variations in the results of certain other tests were so striking that definite conclusions could be drawn.

One possible source of variation was the use of alternative questions in several of the tests. When an entire school system is examined, and the children learn that they will all be tested, the possibility is always present that they will inform each other of the nature of the tests, and the answers to some of the questions. The alternative questions were used to counteract the influence of this factor.

In the test of detecting absurdities in statements, ten or eleven statements were used, the experimenter choosing the five that he would give the subject. The statements varied greatly in difficulty and the experimenters did not use the same selection throughout the experiment. This test was given by B to 26 girls whose average age was 10.6 years, while D gave the test to 25 girls whose average age was 10.9 years. 65% of the girls examined by B passed the test, while only 36% of D's group passed. The variation between the experimenters might be due to the selection of absurdities of unequal difficulty, or to different criteria in grading the responses. The sources of variation are too large to admit of obtaining any reliable results from this test in correlating it with grade and sex.

75% of the girls to whom B gave the test of reconstructing dissected sentences passed, while only 28% of C's girls passed. The average age of the 26 girls to whom B gave the test was 10.8 years, and the average age of C's subjects 10.5 years. Part of the difference between these two experimenters is due to the

fact that more of B's subjects came from the fifth and sixth school grades. Some variation might have been due to different apparatus, B using cards with the sentences printed on two lines, while C had the sentences typewritten on one line. The sentences used by B were more legible, and, being broken into two lines, it was easier to grasp the individual parts as discrete units. Each experimenter used six sentences of varying difficulty so that some variation might be expected from the selection of the three sentences for the test. Whatever the cause of the discrepancy between the results of the two experimenters, it is obviously impossible to obtain any reliable conclusions concerning the correlation of this ability with age, grade or sex, on account of the presence of so many variable factors.

Three problems were used in the test of making change, 20c — 4c, 25c — 6c and 25c — 9c, the process of subtraction involved in each being of unequal difficulty. Certain variations occurred in the tests of comparing remembered objects and comprehending easy and difficult problem questions. Alternative questions were used in both of these tests, and variations might occur due to the relative severity of the experimenters' judgments in marking the responses passed or failed. None of the tests in which alternative questions were used will be treated in the subsequent discussion of the results.

At the close of the experiment, it appeared that a difference of procedure had existed between A and B in the test of indicating omissions in pictures. A and B both showed the three faces first, and the figure with the arms missing last, according to the standard procedure, but A, if his subjects failed to detect the parts omitted from the faces, would give them another trial after they had detected the missing arms. A gave this test to 51 boys and 33 girls, B to 30 boys and 30 girls, his subjects averaging about a year and a half above those of A. The test was passed by 76% of A's boys and 97% of A's girls, but by only 60% of B's boys and 63% of B's girls, showing that the difference of procedure had a most striking effect on the results. It is interesting to note what the effect of a difference of this magnitude would mean if the material from this test were used

as a basis of assigning it to the proper "age group" in the scale. If a test is to be considered normal for a given age if it is passed by 75% of the non-selected school children of that age, the test of indicating omissions in pictures would be a "six year" test for A, and an "eight year" test for B. The data from this test will not be treated in the subsequent discussion.

In the analysis of the results of the definitions test, it was found that certain differences existed between A, B and C in scoring the responses of the subjects as superior to use. No estimates could be made concerning D, for he did not record the actual responses. B, C and D gave this test to approximately the same range of subjects, averaging about 9 years. The corrected results of B and C show, in all, 28% of their subjects giving definitions superior to use, while 65% of D's subjects pass this test. Obviously D was very much more lenient than B and C.

The influence of the personal equation may or may not be present in the remaining tests. In the opinion of the writer it is not present to any marked degree. The data of the four experimenters were treated in several ways, and in none of these was it possible to demonstrate this influence. The writer's opinion, however, is more or less certain according to the test. The tests of repeating digits might show a slight difference between C's results and those of the other experimenters, a difference which could be explained by reference to the rate at which the digits were spoken. The results of experimenter D are slightly lower than those of the other experimenters in the tests of naming 60 words in three minutes and naming rhymes. Whether these differences are real or not, the writer does not know. The data from these tests are included in the subsequent study.

In the subsequent treatment of the results in terms of grade and sex, the material from the following tests will be treated.

VI-2 and IX-2, Defining in terms of use and in terms superior to use.
VII-1, Counting 13 pennies.
VII-2, Describing pictures.
VII-4, Copying diamond.
VII-5, Naming four colors.

VIII-2, Counting backward from 20 to 0.
VIII-3, Enumerating the days of the week.
VIII-4, Counting stamps (three singles and two doubles).
VIII-5, X-3 and XII-1, Repeating 5, 6 and 7 digits.
IX-3, Naming the day and date.
IX-4, Enumerating the months.
X-1, Naming the pieces of money.
X-2, Drawing designs from memory.
X-5 and XI-2, Constructing a sentence, containing one or two ideas from three given words.
XI-3, Giving 60 words in three minutes.
XI-4, Giving rhymes with "day", "mill" and "spring".

The treatment of the results of the definitions test will be confined to the recorded and corrected definitions of A, B and C. The results from the diamond test are based on the scoring system outlined, the passing mark being the fourth group unless otherwise indicated. The arbitrary point system of scoring the design test is used in the subsequent calculations, the passing mark, unless otherwise noted, being 15 points. The corrected scoring of the sentence tests will be used.

The foregoing study of the effect of the personal equation shows conclusively that in certain tests this influence is present to a very marked degree. The errors involved may be traced to three sources, to the apparatus used, to the technique of the experimenters in giving the tests, and to the experimenter's observation in marking the test passed or failed.

The error due to apparatus may result from a variation in the material itself, or from the calibration of different sorts of material as equal in difficulty, e. g.—alternative questions. The variation in the material used by B and C in the test of reconstructing dissected sentences illustrates the error due to defect in the material. The writer has seen apparatus for the line suggestion test in use, in which the last three pairs of lines were actually unequal, the difference between the pairs being above the threshold of discrimination. The subject with good discrimination will invariably fail this test when this faulty apparatus is used.

The error due to the use of alternative questions is more

common and therefore has more practical significance than defects in the material itself. There is a strong temptation for an experimenter, who believes a certain question to be unfair, to substitute another which seems to him to be of the same difficulty. In the study of the Trenton results, which will follow, it will be shown that the different questions included under the same test are not of the same difficulty. The question, "What would you do if you were delayed in going to school?" was passed by practically none of the normal children of 12, 13 and 14. If this question is changed to Goddard's (28) interpretation, "What ought one to do if he is afraid he'll be late for school?", the test is easily within reach of the 12 year children. The difficulty in the first test is caused by the word "delayed". Changing the structure of the test changes its nature completely. In this connection it is to be regretted that Town (72) in the appendix of her translation of Binet's 1911 scale, has changed the wording of some of the tests from that in the actual body of the translation. For example, the question "What would you do before taking part in an important affair?" (page 47) is changed to "Before taking part in something very important, what would you do?" (page 78), and "Why is a bad action done when one is angry, more excusable than the same action when one is not angry?" (page 47), becomes "Why do we more easily pardon a bad act done in anger than a bad act done without anger?", (page 79). The meaning is the same but the wording different, and in many cases success or failure in a test depends on the interpretation of a single word. If an experimenter using Town's translation were allowed to select his questions from the actual translation or the appendix indiscriminately, variations would, in all probability, result. The general proposition that there is no such thing as an alternative question, i. e. a question involving the same mental processes and having the same difficulty as another, could very easily be maintained. To avoid this error experimenters should adhere strictly to one wording and should never be allowed to substitute one question for another.

An example of the influence due to differences of the tech-

nique of the experimenters in giving the tests is afforded by the test of detecting omissions in pictures. This test is a "six year" test for A and an "eight year" test for B. Differences in procedure make it very difficult if not impossible to compare the results of one investigator with those of another. To eliminate this error, very careful and minute instructions should be published for the giving of each test. No edition of the Binet-Simon scale is entirely satisfactory in this particular.

Examples of errors due to the observation of the experimenters are afforded by the tests of copying a diamond and defining in terms superior to use. Errors due to observation may be avoided or minimized by increasing the number of grades of response with which the particular response in question may be compared. This principle is followed by Yerkes (82) in the arrangement of the Point Scale. In the diamond test, for example, Yerkes allows three grades of response while Binet allows but two—plus or minus. The accuracy of any measure increases with the number of gradations on the measuring scale, and the significance of the error of observation is diminished by decreasing the chances of wide displacement. In the tests in which a definite question is put to the subject, uniformity of scoring may be obtained by an accurate and painstaking cataloguing, and a subsequent classification and weighting of all the responses of a large number of subjects to each question. If the responses to a free association test may be classified into a relatively small number of groups, then the responses to a restricted association test could be classified into a much smaller number of groups. A sufficiently large number of responses will include practically all possible responses. In this way the chances of the error due to observation are diminished, while the adoption of a point system of scoring will minimize the effect of any errors that might be made.

The differences between the experimenters in this study are large enough to demonstrate the influence of the personal equation. Scientific procedure demands that the investigator who studies the results of the individual tests for the purpose of analysing the factors involved or for obtaining age norms should

demonstrate that the effect of the personal equation is not present in the results treated. The burden of proof should be on the person who maintains that the influence is not present. Negative results concerning the influence of the personal equation that are based on the method of comparing the total scores or "mental ages" of different experimenters should not be taken as conclusive, inasmuch as the experimenters may deviate in one direction in one test, and in the opposite direction in another, so that in a total score these deviations might equalize. In a study of this sort made on the basis of "mental ages," which has previously been reported, the writer (14) found no deviations between B. C and D, while deviations between these three experimenters do appear in the more detailed study of the individual tests. Studies of the individual tests can have no claim to reliability unless the personal equation has been eliminated.

The importance of the personal equation as a source of error in making diagnoses on the basis of the "mental age" of the subject is universally recognized by psychologists and almost universally ignored by medical men, field workers, school teachers and others who have had no experience in making mental measurements. Among psychologists there are two opinions concerning the solution of the difficulty arising from this source, the first, that of making certain allowances for the inexpert examiners or establishing limits within which their opinions are valid, the second, that of removing the scale from their hands entirely.

Doll (22) in discussing criticisms of the Binet scale on the ground that diagnoses of normality and feeble-mindedness are made by inexpert examiners urges "that those who are capable of doing good Binet testing of the mechanical sort without being clinical psychologists should report the findings of their examinations of children or groups in tables of related chronological and mental ages and not in terms of normality or abnormality. In their reports they can say with a high degree of certainty that those children who show an intellectual retardation of more than 3 years are feeble-minded, but they should not say that those who test less than 3 years retarded are backward or normal. In

the lesser degrees of retardation only the expert is capable of evaluating the details of a Binet test with any finality as to either diagnosis or prognosis." (page 607).

Doll also points out that Binet examiners who have worked in institutions give very reliable diagnosis, for they intuitively sense distinctions which inexpert laymen do not see. When the responsibility for the diagnosis is placed on the examiner in this way, the scale it treated as a qualitative instrument. This standpoint is quite different from that in which certain allowances are made for all inexpert examiners and the quantitative character of the scale preserved. Goddard (31) in a study of the personal equation based on re-testings of normal and feeble-minded individuals fixes the quantitative limits somewhat higher. "In all cases where a child tests four or more years behind his age, there is little danger of error in considering him feeble-minded, even though the test was made by a person who was not highly expert, provided such a person is able to use the test with reasonable intelligence. With the borderline cases, those who are two or three years backward, the best expert should be employed in the testing." (pages 76-77).

As early as 1910, before the scale had received very extensive application, Huey (35) took the stand that inexpert examiners should not use the scale. In discussing this point he said, "I would urge that these Binet tests must be used with judgment and trained intelligence, or they will certainly bring themselves and their authors into undeserved disrepute.—Results can be considered valid only when the tests are made by an experienced psychologist who has familiarized himself with Binet's directions, or by other competent persons who apply the tests under the direction and supervision of such a psychologist." (page 444).

Three years later, in referring to the reports that the medical inspectors in Pittsburgh were to take over the Binet testing in the schools, Whipple (78) says, "And we can only express our hopes that these reports are unfounded, or that at least those in authority may be led to understand that for a person, whoever he may be, without extensive psychological training to attempt to diagnose the precise mental status of a school child is about as

absurd as for a mere psychologist to attempt to diagnose incipient tuberculosis or any other obscure pathological condition." (page 302). The same position is taken by Whipple (77) in another editorial. "We have no quarrel with the use of the scale in the public school: properly used, it is of direct and practical value; but improperly used, it will become a farce which can but bring discredit upon psychology and retard the movement for its application to educational practise." (Page 119).

In defense of this position, Whipple calls attention to an error inherent in the procedure of all inexpert examiners. "There is nothing about the conduct of the Binet-Simon tests that is intrinsically difficult, yet there is a source of error inherent in the use of any psychological procedure, which, as experience shows, is surmountable only by drill in psychological experimentation. I refer to the difficulty of following directions. No one who has drilled students in the laboratory has failed to be struck with the impossibility of laying down fool-proof directions for the conduct by an amateur of a psychological test." (Page 119).

Kuhlmann (43) agrees with Whipple in this position. "The untrained examiner meets difficulties because he lacks the following: (a) Familiarity with the directions for giving the tests. (b) Familiarity with the rules for interpreting the responses of the children. (c) Ability to adapt the procedure in testing in special instances for which directions can not be given. (d) Ability to interpret responses in special instances for which rules can not be given. (e) Ability to adapt himself in attitude to the mental levels of children of different ages so as to obtain the best efforts from the child in each case. (f) General appreciation of the absolute necessity of adhering strictly to all rules of testing, and of careful, painstaking work. These deficiencies are of quite different degrees of importance. The last is, on the whole, the most serious and most frequent, and can be remedied only by extended laboratory training." (Pp. 255 and 256). In regard to the quantitative allowance that must be made for inexpert examiners, Kuhlmann's article affords the following, "The amount of error made by an examiner because of his lack of training seldom equals two years in the mental age; in the majority of cases it is less than one year." (Page 256).

IV. GRADE CORRELATIONS

The correlation between intelligence, as measured by the Binet scale, and school performance, as measured by age and grade standing, has been worked out by various investigators. In all cases intelligence was measured by the "mental age" or total score of the Binet tests, and pedagogical age by assuming that all children begin school at a certain age and should therefore be in certain grades at certain ages. Stern (62) has reviewed the work of Goddard (30), Binet (4), and Bobertag (10) in this field, with the general conclusion that the correlation is only moderately high. The number of children showing mental advance is in excess of those showing pedagogical advance, but very rarely do children showing pedagogical retardation show mental advance. The correlation is one-sided in that "inference from school performance to mental ability is safer than from mental ability to school performance." (Page 61). Stern accounts for the discrepancies on the ground that "performance in the school depends not only upon intelligence, but also upon certain other and quite different factors." (Page 63). These factors are strength of memory which plays a large part in school performance but correlates only to a moderate degree with intelligence, and other factors that have nothing to do with intellect but belong largely in the domain of the will—"the degree and duration of attention, industry and conscientiousness, sense of duty and capacity to fit into the social group." (Page 63). Stern concludes that "the lack of agreement between tests of intelligence and school performance is really calculated to increase our confidence in the psychological test-methods," (Page 64) that absolute correlation is not to be desired since that would mean that the tests were testing school performance only, and that the measure of intellectual ability was the school performance itself, the tests being superfluous.

More recently, Schmitt (57) has reviewed the work of God-

dard, Terman and Childs (66) and Dougherty (23) in correlating intelligence, as measured by the Binet scale, and school performance, and reaches conclusions quite opposite to those of Stern. The following quotations from Schmitt's monograph explain her view point. "Further doubt is cast upon the accuracy of the tests by the fact that judgments arrived at through their application do not coincide with that of the school concerning the same subjects." (Page 57). Concerning this lack of correlation Schmitt writes "The Binet tests, therefore, while professing to test native ability are concerned very little with the education which all normal children have the native ability to acquire, and which is of much importance in civilized life." (Page 60). To the investigations cited Schmitt has added one of her own, in which the lack of correspondence between the Binet "mental age" and school grade is shown.

The writer is of the opinion that the method of correlating school performance with "mental age" fails to demonstrate either the adequacy of the Binet tests according to Stern, or the complete inadequacy of the tests according to Schmitt. For the demonstration of this point Schmitt's investigation may be discussed, inasmuch as it shows the most striking deviations between the measures of the two performances. Schmitt applied Binet's 1911 scale (Town's translations with modifications) to 150 children of superior social status. The following quotations indicate the status of the subjects. "The children who served as subjects for the tests comprised the Kindergarten and first six grades of a private school in Chicago." "They were the children of the professional class mainly. A few were children of successful business men who sought the best obtainable type of education for their children." (Page 2). The tests were applied at the close of an examination with the Healy-Fernald tests under rather unfavorable conditions as indicated by the following quotations,—"In the conduct of the two sets of tests the Binet-Simon tests were reserved for the last. By the time they were reached the child had been doing tests for an hour or more. In some cases there was too much restlessness and fatigue to

carry the child as far as the majority of his comrades in his grade were able to go and the tests were then discontinued." (Page 68 and 69).

The tests in the various age groups given to each grade were as follows,—Kindergarten, tests for V, VI, VII, VIII and IX years; Grade I, tests for V, VI, VII, VIII, IX, X and XII years; Grade II, tests for VI, VII, VIII, IX, X, and XII years. Grades III and IV, tests for VIII, IX, X and XII years; Grade V, tests for IX, X, XII and XV years; Grade VI, tests for XII and XV years. The "Adult" tests were also given to Grade VI as a class-room test.

Schmitt compared three measures, chronological age, school grade age and "mental age". The "mental age", in case a subject passed all tests in one group and failed one or more in a lower group, could be reckoned from two basal ages, these alternative rating being included by Schmitt. The summary of the results is as follows,— Comparing the Binet age to the chronological age, 14 (or 20)% are retarded, 26 (or 24)% are normal, and 58 (or 54)% are advanced. Comparing the school grade to the chronological age, (using 5 to 6.5 years as the normal age for the Kindergarten, 6.5 to 7.5 for Grade I etc.) 38% are retarded, 56% are normal and 4% are advanced. Comparing the Binet age to the school grade age, 2 (or 4)% are retarded, 25 (or 35)% are normal and 72 (or 60)% are advanced. The essential discrepancies are indicated by Schmitt by the following,— "Where the school grading shows 4% advanced over the normal for the chronological age, the Binet grading shows 58% over the chronological age and 72% over the age normal to the school grade." (Page 80.) The discrepancies thus indicated, although much larger than those of other investigators, agree with the general trend of results in that more children are shown to be advanced according to the Binet mental age than according to the school grade age. The results disagree with those of other investigators in finding a higher per cent. advanced by Binet age compared to chronological age.

The inadequacy of the methods employed in the investigations of Schmitt and others is seen when the measures are separately

studied. The use of the normal grade age as a measure of scholastic ability is false inasmuch as it rests on the assumption that all children enter school at a certain age, which is not the case. The measure of scholastic ability is the measure of the child's reaction to the subject matter of the grades, and that measure may be expressed only in the fact of promotion, non-promotion or (very rarely) double promotion, in other words, it may be expressed only in the relation of grade to the length of time in school. Furthermore, the two measures of scholastic ability, the age in grade method, and the grade progress method, are measures of an historically past performance not of present possibilities, and the true measure of an ability must indicate potential ability.

As measures of scholastic ability in terms of actual reaction, these measures present a distribution of general ability that is skewed toward the lower end, or in the direction of no ability. If a child enters school late, he presents a picture of retardation according to the age and grade method, while through any number of causes independent of intellectual ability, a child may present a retardation of at least a year according to either method. The possibilities for advancement are not as great, however, for advancement means forcing a child through a mass of subject matter, a process which the school is generally unwilling to undertake and the parent is generally unwilling to sanction. The school therefore presents a picture of ability in which promotion is normal, and non-promotion far more frequent than advance. If general ability is to be considered as distributed over any sort of a frequency surface, that surface will not take the form presented by the school measure in which the modal ability is almost completely the upper limit.

The measure of "mental age" has been shown to be one which varies from one chronological age to another in the form of its distribution. Normal children of 6 or 7 test over age, while those of 11 and 12 test under age. This abnormal distribution is due to two facts. In the first place, the tests in the younger years are too easy and those in the higher years are too difficult.

In the second place, the younger children have a wider range of tests beyond their average ability, so that exceptional subjects may display exceptional ability in a manner that is impossible if ability is measured by school progress, while older children have only a few tests within their range, the picture of advancement being excluded as in the measure of school ability. If the mental ages of a run of subjects of different chronological ages are combined, the frequency surface is normal, the error of the extremities balancing.

The investigators who have compared "mental age" with grade age, have compared two distributions, one of which is markedly skewed, the other normal, but false. The resulting finding of mental advance in excess of pedagogical advance has significance only insofar as it shows that a measure of general ability that will admit of exceptionally high performance is a better measure than one that precludes the possibility of such performance. The only significant finding is that pupils who show marked retardation in school rarely if ever show mental advance.

Applying the foregoing discussion to Schmitt's results in particular, all that has been said concerning the inadequacy of the age in grade method applies to her results. The age for entering school being 5, none of the subjects in the Kindergarten could be advanced, while those who entered late would be retarded. It is difficult to see how these young children would be able to make up their work in such a way as to show advance during the first two or three school years. The normal age for the sixth grade is from 11.5 to 12.5 years. Inasmuch as no grades were tested above VI, none of the 37 subjects from 11.5 to 14.5 could show an advance, and all of the 19 subjects from 12.5 to 14.5 would necessarily show retardation. Schmitt's results differ from those of other investigators in finding more subjects advanced according to Binet age in relation to chronological age. This deviation is probably due to the fact that she examined a superior selection of subjects, and to the fact that the XV year and "Adult" tests were used, so that the older subjects, who in general fall below their chronological age, had an opportunity to

better their scores. The discrepancy shown by Schmitt between school standing and the Binet tests does not demonstrate the inadequacy of the tests.

The final demonstration of a correlation between the Binet scale and school grade, rests not in comparing the total score or "mental age" with school grade, for that is susceptible to the errors of over-estimation and under-estimation according to varying chronological age, but in comparing the results of subjects in each grade on the individual tests. The tests may vary in their correlation with grade. Inasmuch as there is a general growth in age with grade, and a corresponding growth of intelligence with age, a test, in order to be an adequate test of intelligence, must show a correlation with grade. If the correlation is too high, however, the value of the individual test is in question for it would then be testing, not intelligence, but grade training. This criterion was actually used, though not stated, by Binet in his discussion of the results of Decroly and Degand (19), and in his revision of the 1908 scale, in which many of the tests that he considered to relate to school training were eliminated.

Studies of the individual tests in the light of school grade are not available. Decroly and Degand published in 1910 the results of an investigation on 45 children in a Brussels school, similar in character to that studied by Schmitt in Chicago. Binet discussed these results and those of other minor investigations in the Paris schools in considering the effect of environment on the results of the tests. Although he referred to school training as a factor, and classified the tests in which Decroly and Degand's subjects were superior, he gave no quantitative demonstration of the effect of this factor. The results of Decroly and Degand are based on too few subjects to admit of quantitative treatment. Chotzen (18) studied the tests by comparing the performance of feeble-minded individuals of the same mental age but of different chronological age. Although this method shows the effect of environment and maturity on feeble-minded individuals, it does not bear directly on the factor of school training. The foregoing

investigations will be discussed in this chapter only in their relation to the results of the particular tests. Schmitt, in her monograph, published tables showing the reaction of each subject in each grade to each test, the tables being discussed in the text. Although it was not Schmitt's purpose to determine the correlations between the various tests and grade, her data are available for a study of this sort, and the writer has taken the liberty of figuring them in this light, indicating at the same time Schmitt's interpretation of the grade factor, contained in the accompanying text. These data will be compared with the results of the Princeton investigation.

422 subjects of this investigation were distributed in the kindergarten, first six regular grades, minus grades and the special class of the Princeton Model School. 301 of the subjects (161 boys and 140 girls) were in the kindergarten and first six regular grades. The data obtained from the examination of these 301 subjects were classified according to the grade in which the subjects were found, and the percentage that the subjects of each grade passed each test was calculated.

Only those tests were studied which showed themselves to be free from the influence of the personal equation of the four experimenters. The elimination of the unrecorded results of the definitions test left a number of cases too small to be studied. To avoid the influence of the error due to incomplete data, the writer has calculated the percentage from only those tests that were given from 75% to 100% of the possible number of times. The data from the tests of repeating 5, 6 and 7 digits have been combined into one weighted measure. The procedure of the experimenters in giving these tests was to start within the subject's range and continue till he failed. If 5 digits were successfully repeated, 6 were given, and if these were passed, 7 were given. The results have been combined into one measure for the sake of simplicity, 1 point being allowed for the successful repetition of 5 digits, 2 points for 6 digits and 3 points for 7 digits, the weighting being roughly in accordance with the weighting in Goddard's scale, the tests being in the age groups

VIII, X and XII respectively. The measure of the ability of a group to repeat digits is the per cent. that the number of points scored is of the number of points possible (i.e. 6 times the number of subjects in the group).

The number of subjects in each grade (boys and girls shown separately) the average age of the subjects in each grade, together with the mean variation from the average are shown in Table 3.

TABLE 3

Number of Boys and Girls in Each Grade, and the Average Age of All Subjects in Each Grade.

Grade	Number of Boys	Number of Girls	Total No. of Subjects	Average Age	Mean Variation
Kindergarten....	20	12	32	5.64 years	0.46 years
Grade I.....	27	24	51	7.05 "	0.50 "
Grade II.....	16	24	40	8.16 "	0.65 "
Grade III.....	21	24	45	9.31 "	0.75 "
Grade IV.....	20	15	35	10.46 "	0.91 "
Grade V.....	24	25	49	11.71 "	0.99 "
Grade VI.....	33	16	49	12.81 "	1.06 "

The above table shows an increase of a year or more (actually from 1.10 years to 1.41 years) in the average age of the subjects in each grade. From this it is reasonable to expect that there is a general growth in intelligence correlating with this increase in age, or, in other words, to expect a correlation between the results of the individual tests and the grade in which the performance occurred. If the correlation is too high, it will indicate a dependence of that particular test on the subject matter of the grade. In Table 4 are shown the percentages that the subjects in each grade passed each test. The notes referred to in the margin contain the proportions passed for all other subjects for whom the percentages are not given, the percentages being given only for those groups to whom the tests were given from 75% to 100% of the possible number of times.

A study of Table 4 shows that the tests in general correlate with grade. The combined score of the test of repeating digits, for example, shows a growth from 6% to 78%, more rapid in the first three grades than in the last four. The tests vary in

TABLE 4

Percentage that Subjects in Each Grade Passed Each Test. 301 Subjects.

Test	Grades K	I	II	III	IV	V	VI	
VII-1, 13 pennies	72	96	100					Note 1
VII-2, Pictures	69	96	94					Note 2
VII-4, Diamond	46	75	88					Note 3
VII-5, Colors	72	90	97					Note 1
VIII-2, 20 to 0		9	53	80				Note 4
VIII-4, Stamps		13	50	78				Note 5
All digits, (combined)....	6	21	42	51	55	78	75	
VIII-3, Days of week....	16	45	90	100				Note 6
IX-3, Date		5	35	96	100			Note 7
IX-4, Months			28	84	90			Note 8
X-I, Money			20	36	57	82		Note 9
X-2, Designs				21	37	42	66	Note 10
X-5, Sentence (2 ideas)..				67	89	88	98	Note 11
XI-2, Sentence (1 idea)....				22	46	51	74	Note 12
XI-3, 60 words					63	63	87	Note 13
XI-4, Rhymes					67	63	76	Note 14

Note 1. Counting 13 pennies and naming colors given 20 times above II. Not failed.
Note 2. Describing pictures given 21 times above II. Not failed.
Note 3. Copying diamond given 25 times above II. Not failed.
Note 4. Counting from 20 to 0 given 18 times in K. Not passed. Given 31 times above III. Failed once.
Note 5. Counting stamps given 15 times in K. Not passed. Given 35 times above III. Failed 3 times.
Note 6. Naming days of week. Given 32 times above III. Not failed.
Note 7. Giving day and date given 5 times in K. Not passed. Given 56 times above IV. Not failed.
Note 8. Naming months. Given 26 times below II. Passed twice. Given 44 times above IV. Failed twice.
Note 9. Naming money. Given 26 times below II. Passed 3 times. Given 28 times in VI. Failed twice.
Note 10. Copying designs given 33 times below III. Passed 5 times.
Note 11. Sentence (2 ideas) given 32 times below III. Passed 12 times.
Note 12. Sentence (1 idea) given 32 times below III. Passed 4 times.
Note 13. Giving 60 words given 53 times below IV. Passed 19 times.
Note 14. Giving rhymes given 42 times below IV. Passed 26 times.

the number of grades taken to reach their maximum. The test of naming the day and date, for example, is failed by all subjects in the kindergarten, 95% of Grade I and 65% of Grade II, while only 4% of the subjects in Grade III and none of those in the

higher grades fail it. A sudden increase occurs between Grades II and III showing possibly the influence of grade training. The tests vary considerably in the degree of their correlation. An easily obtained measure of the degree of correlation is that of comparing the magnitude of the increases from grade to grade. For example, there is an increase of 61% (96%—35%) from Grade II to Grade III in the ability to pass the test of giving the day and date, and an increase of 16% (36—20%) between the same grades in the test of naming the pieces of money. The former test correlates higher with the influence of grade in this particular case than the latter.

In this manner the percentage difference between the performance of the subjects in each grade and that of the subjects in the preceding grade was obtained. All the increases or decreases in ability from one grade to another were thus obtained, these values serving as measures of the amount of correlation between the tests and the grades. 42 differences between the performance of the subjects in any grade and those of the next succeeding grade were thus obtained. In 4 cases there were actual decreases of 1, 2, 3 and 4% which were not significant. The difference ranged from —4% to +61%, the median being +19.5% (Q=16.25%). Some of the differences between the grades might be due to the chance superiority of a particular grade. To overcome this chance variation, and to furnish another index of the growth of the various abilities, the differences were calculated by steps of two grades, i.e., subtracting the performance of the kindergarten from the second grade, the first from the third, etc. In this way, 26 differences were obtained varying from +9% to +91%, the median being +29% (Q=18%).

Some of the differences noted are undoubtedly high enough to warrant the assumption of the effect of grade training on the tests. Just what tests show this effect is probably a matter of opinion. Allowance must be made for the growth of an ability independent of training. 25% of the highest increases from one grade to another were selected as being worthy of special consideration at least. A larger increase must be allowed be-

tween two grades. Those differences were considered worthy of special consideration that exceeded twice the value of the median of the one-grade differences or 39%. This manner of selecting the largest differences is quite arbitrary, but is justified by the outcome, for the tests that show the most significant increases according to this method show those increases in more than one step, so that the evidence is concentrated against a very few tests. In this way the significant values outweigh the less significant values and fair allowance is made for growth from one grade to another.

The following list includes the tests showing the greatest increases by one-grade and two-grade steps, together with the magnitude of the increases and the grades between which they occur.

One-grade steps. 25% of largest increases.	Two-grade steps. Increases greater than 39%.
+61% Date, II to III	+91% Date, I to III
+56% Months, II to III	+74% Days, K to II
	+71% 20 to 0, I to III
+45% Days, I to II	
+44% 20 to 0, I to II	+65% Stamps, I to III
	+65% Date, II to IV
+37% Stamps, I to II	+62% Months, II to IV
+30% Date, I to II	
	+55% Days, I to III
+29% Diamond, K to I	
+29% Days, K to I	+46% Money, III to IV
+28% Stamps, II to III	+42% Diamond, K to II
+27% 20 to 0, II to III	
+27% Pictures, K to I	

The above lists of increases are confined to but 8 tests. In all, there were 16 tests studied. According to the method of selecting the significant increases, 20 such values actually appeared. In this manner the evidence combines against a very few tests. Some tests appear in both lists and more than once in the same list. The most striking growth with grade is shown in the tests of giving the day and date, naming the months, nam-

ing the days of the week, counting from 20 to 0 and counting stamps. The tests of copying the diamond, describing pictures and naming money may or may not show this influence. The evidence is strongest in the case of the diamond test since that appears in both lists.

The foregoing method of selecting those tests which correlate with grade to such an extent as to indicate the influence of grade training is not conclusive, owing to the fact that there is also an increase in age from grade to grade. If a test showed a very rapid growth with age, and those ages fell for the most part in certain grades, then those grades would show an increase which might be wrongly assumed to be due to training. The tests of counting from 20 to 0 is a case in point. Yerkes (82) in Table 32, page 125, gives the percentage values for each test in the Point Scale, for English speaking boys and girls of each age. The test, of the twenty one tests included, that shows the most marked increase with age is that of counting backward, the values being as follows,— age 4=0%; age 5=3.5%; age 6=23.7%; age 7=45.7%; age 8=72.2%; age 9=96%; the values for ages above 9 being 97% or higher.

The age in grade distribution of the 301 subjects in this investigation is given in Table 5.

TABLE 5

Distribution of Subjects in Each Grade according to Chronological Age.

	Grades							
Age	K	I	II	III	IV	V	VI	Total
4	4							4
5	17							17
6	11	28	2					41
7		18	17	2	1			38
8		4	15	18	1			38
9			5	13	11			29
10		1		10	14	18	1	44
11			1	2	3	16	16	38
12					5	8	11	24
13						4	12	16
14						2	5	7
15						1	3	4
16							1	1
Total	32	51	40	45	35	49	49	301

The rapid growth of the ability in counting from 20 to 0, according to the method of comparing the subjects in each grade, was from 9% in Grade I to 80% in Grade III. From Table 5 it may be seen that practically all, (89%), of the chronological ages in Grades I, II and III were distributed in the ages 6, 7, 8 and 9, a chronological range coinciding with that in which Yerkes' results show the ability to develop. The growth of this ability might be due then either to age or to grade. For this reason, to arrive at any final conclusion, it is necessary to compare the subjects of the same age but in different grades. The treatment of the Princeton results according to this method follows, but the analysis of the data in this manner can have no great reliability owing to the small number of subjects in each group. The number of subjects in each group, (boys and girls shown separately), the average age and mean variation from this average are shown in Table 6.

TABLE 6

Number of Boys and Girls of Similar Ages in Different Grades, and the Average Age of the Subjects of Similar Ages in Each Grade.

Grade	Age	Number of Boys	Number of Girls	Total no. of Subjects	Average Age	Mean Variation
Kindergarten.	5	11	6	17	5.48	0.20
Kindergarten.	6	8	3	11	6.26	0.21
Grade I	6	14	14	28	6.59	0.17
Grade I	7	9	9	18	7.36	0.22
Grade II	7	7	10	17	7.56	0.24
Grade II	8	6	9	15	8.39	0.24
Grade III ...	8	8	10	18	8.60	0.22
Grade III ...	9	5	8	13	9.43	0.16
Grade IV	9	5	6	11	9.65	0.13
Grade IV	10	10	4	14	10.39	0.30
Grade V	10	7	11	18	10.54	0.25
Grade V	11	10	6	16	11.54	0.22
Grade VI ...	11	10	6	16	11.53	0.26
Grade VI	12	6	5	11	12.52	0.14

All chronological ages were computed in tenths of a year, so that a variation in age from 0.1 yr. to 0.9 yr. is possible within

TABLE 7

Actual percentage that each test was passed by the subjects of each age in each grade. 223 subjects.

AGE	Age 5	Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12
GRADE	K	K	I	I	II	II	III	III	IV	IV	V	V	VI	VI
NO. OF SUBJECTS	17	11	28	18	17	15	18	13	11	14	18	16	16	11
Counting 13 pennies	71	91	96	94	100	100	Note 1							
Describing pictures	65	82	96	100	100	86	Note 1							
Copying diamond	50*	45	64	89	93	92	Note 2							
Naming colors	71	82	89	100	100	100	Note 1							
Counting from 20 to 0			16	0	63	43	83*	Note 3						
Counting stamps			13	12	29	60	85*	Note 4						
Repeating digits, (all)	6	6	21	19	42	46	47	51	65	48	79	83	76	86
Naming days of week	0	40	50	50	94	80	100*	Note 5						
Giving the day and date					29	20	88	100	100	100	Note 6			
Naming months					25	27	88	85	100	Note 7				
Naming pieces of money						31	35	31	55	64	83	83	Note 8	
Copying designs							31	31	27	43	41	57	Note 9	
3 words in sentence. 2 ideas				Note 10			89	62	91	93	94	75	93	100
3 words in sentence. 1 idea				Note 11			33	8	45	50	61	50	80	80
60 words in 3 minutes					Note 12				44	77	50	88	93	80
Giving rhymes					Note 13					92	56	69	87	100

*The test of copying the diamond was given 59% of the possible number of times in K-5, counting from 20 to 0 and giving days of week, 66% in III-8, and counting stamps 72% in III-8. All other percentages are based on tests given from 75 to 100% of the possible number of times.

Note 1. Tests of counting 13 pennies, describing pictures and naming colors each given 12 times above II-8. No failures.
Note 2. Copying diamond given 15 times above II-8. No failures.
Note 3. Counting from 20 to 0 given 16 times below I-6. Not passed. Given 31 times above III-8. Failed 4 times.
Note 4. Counting stamps given 14 times below I-6. Not passed. Given 32 times above III-8. Failed 4 times.
Note 5. Giving days of week given 32 times above III-8. No failures.
Note 6. Giving date given 39 times below II-7. Passed twice. Given 36 times above IV-10. No failures.
Note 7. Naming months given 24 times below II-7. Passed twice. Given 37 times above IV-9. Failed 4 times.
Note 8. Naming pieces of money given 35 times below II-8. Passed 4 times. Given 14 times above V-11. Failed twice.
Note 9. Copying designs given 26 times below III-8. Passed 5 times. Given 15 times above V-11. Failed 6 times.
Note 10. Three words in sentence, 2 ideas, given 24 times below III-8. Passed 9 times.
Note 11. Sentence, 1 idea, given same as 2. Passed 3 times.
Note 12. 60 words in 3 minutes given 41 times below IV-9. Passed 10 times.
Note 13. Giving rhymes given 37 times below IV-10. Passed 25 times.

each age group. That the subjects of the "same" age but in different grades are not exactly the same is shown in Table 6. The subjects of each age in the higher grades average from 0.01 yr. to 0.33 yr. different, with an average superiority of 0.19 yr. This difference, however, is about one fourth that between the subjects or different ages in the same grades, and may be called the same for practical purposes. For convenience, the groups will be referred to as K-5, II-7 etc., the first member referring to the grade, the second to the age. K-5 would mean the group of 5 year children in the kindergarten, II-7, the 7 year subjects in Grade II, etc. The actual per cent, that the subjects in each group passed each test was calculated and is shown in Table 7. Unless otherwise noted, the percentages are based on tests given 75% to 100% of the possible number of times.

Some of the groups from which results were obtained are too small to have great reliability, but the method is at least suggestive. The results of 14 groups are given. It is possible then to compare the results of subjects of 6 ages, (6, 7, 8, 9, 10 and 11), that are in different grades, and also to compare sub-

jects in all seven grades that are of different ages, and in this way to determine whether the dominating factor in the growth of any ability is that of grade or age. The reliability of the method rests only on its connection with that of the first method employed.

In answer to the question of whether the growth of ability in the test of counting from 20 to 0 is due to age or grade, a question which was unanswered by the first method, we may turn to the results shown in Table 7 in which the subjects of each age in each grade are shown. The test of counting from 20 to 0 was not passed by any of the 5 and 6 year subjects in the kindergarten. Comparing first the subjects of different ages in the same grade, the 7 year subjects in Grade 1 are 16% lower than the 6 year subjects in that grade, and the 8 year subjects in Grade II are 20% lower than the 7 year subjects in the same grade, the older subjects making a lower record in each case. Comparing the performance of the subjects of the same age but in different grades, the 7 year subjects in Grade II are 63% ahead of the subjects of the same age in Grade I, while the 8 year subjects are 40% [1] ahead of the subjects of the same age in Grade II. Allowing for the retrogression of the older subjects in each group, i.e. assuming that they should have done equally as well as the younger subjects in the same grade, the groups in Grades II and III are still 47% and 20% ahead of the subjects in the grades lower. The growth of ability in this test would therefore appear to be due to grade training.

A rapid growth of ability in the test of counting stamps occurred between Grades I and III (37% I-II+28% II-III=65% I-III), according to the first method, so that the same question arises as in the test of counting from 20 to 0. The test was not passed below group I-6. No growth with age is shown between

[1] This test was given to but 66% of the subjects in III-8, the experimenters assuming that the other 34% would pass. The score given, 85%, therefore represents the ability of the lowest selection of III-8 subjects, or the most conservative estimate of the ability of the whole group. The same applies to the other tests in III-8 given 66% and 72% of the time. In this way the hypothesis that the tests are not influenced by grade training is given the benefit of the doubt.

I-6 and I-7, but a growth of 31% appears between II-7 and II-8. A growth with grade of 17% is shown from I-7 to II-7 and of 25% from II-8 to III-8. This test shows therefore the operation of the two factors of age and grade training.

The improvement in ability in the tests of counting 13 pennies, describing pictures and naming colors, that was indicated between the kindergarten and Grade I by the first method, would refer to age rather than grade, for a greater increase in each test is indicated between K-5 and K-6 than between K-6 and I-6. Above I-6 these abilities are completely developed. It could be maintained that these tests are so completely within the ability of the groups that the effect of training would not be indicated. The test that is best adapted to show the influence of any factor on a group is one that is well within the ability of the group—the influence of the factor will be obscured if the measure is either too easy or too difficult. The test of copying the diamond is a case in point and one well worth study, for it has been attributed to the effect of training by various authors. All the reproductions of the diamond had been scored according to the arbitrary system outlined in the previous discussion of the personal equation. A control on the factor of difficulty was obtained by raising or lowering the passing mark in this test. The percentage passed was calculated for each group for each of the 5 possible passing marks. The relations indicated in Table 7, where the passing mark is Group IV, were not changed by this process of raising or lowering the passing mark. In all cases the influence of age was shown between groups I-6 and I-7, and the influence of grade shown between groups K-6 and I-6. The test was given to but 59% of the K-5 group, the experimenter assuming that the other 41% would fail, so that the percentages calculated represent the performance of the best selection of K-5 subjects, or, in other words, the benefit of the doubt is given to the hypothesis that the test is influenced by grade training. If the other members of K-5 had failed according to the experimenter's assumption, (and this assumption was quite justified for some had failed to draw the square), 29% of the group would have passed instead of 50%.

The influence of age indicated in this test is as great if not greater than that due to training.

The test of repeating digits, scored by the weighting system previously described, exhibits a slow but uniform progress throughout, the older subjects in each group making records that are about the same or slightly lower than those of the younger subjects in the same grade, an increase showing fairly regularly from grade to grade. The most marked increase in this ability appears between K-6 and I-6, and between I-7 and II-7, possibly indicating that the lack of familiarity with the use of digits in the lowest grades interferes with this test as a measure of auditory memory.

The test of naming the days of the week shows the most marked improvement with age (40%) from K-5 to K-6, practically no improvement (10%), from K-6 to I-6, no improvement from I-6 to I-7, a very marked increase with grade from I-7 to II-7, a drop from II-7 to II-8, group III-8 marking the complete development of the ability. The test would appear to be due to the combined effect of age and grade. The tests of giving the day and date and naming the months are passed only twice in the kindergarten and first grade, by about a quarter of the subjects in II-7 and II-8 without age increase, while the subjects in III-8 shows a most marked increase due to grade. Above III-8 these tests are seldom failed. The test of naming the pieces of money shows a slow growth from 8 to 11, the largest increases appearing from III-9 to IV-9 and from IV-10 to V-10, improvement with grade in each case. Copying the designs from memory shows a growth of 26% from 8 to 11, the development occurring in two age steps, from IV-9 to IV-10 and from V-10 to V-11.

The growth with age cannot be determined in the tests of constructing sentences from three given words, because they were given to too few cases below the third grade. The results do not show whether III-8 is exceptionally high or III-9 exceptionally low. Both tests show decreases in ability from III-8 to III-9 and from V-10 to V-11. The ability in the easier test is well within the range of the third and higher grades, showing, therefore, no

improvement. The improvement in the second test develops from 33% to 80% in three steps, correlating with Grades IV, V and VI in each case. The most vital question, that of determining whether or not the language training in the third grade helps to make the construction of a sentence possible, cannot be determined owing to the lack of material in the second grade. The experimenters' assumptions in not trying the test would indicate this fact, but this is not experiment. The same lack of material makes conclusions in regard to the rhyming test impossible. The performance of IV-10 is exceeded only by VI-11. The test of naming 60 words in three minutes shows two decided increases with age and one decided drop with grade.

The foregoing analysis is based on a number of subjects in each group too small to have any great significance. The general fact of the correlation of the tests with grade remains, and conclusions concerning what tests correlate too highly with training can be answered only by considering both methods of study, and by considering only the largest deviations. The two most striking instances are found in the tests of naming the months and giving the date. These tests undoubtedly relate almost entirely to training. Less striking but equally definite is the relation of the test of counting from 20 to 0 to training. The tests of naming the days of the week and counting stamps show the influence of age to an extent almost as marked as that of grade, so that while the development in these tests is rapid, the grade factor probably exerts only part of the influence. Conclusions concerning the other tests are largely a matter of opinion, and the opinion of the writer has been indicated in the detailed discussion.

A study of the tests in relation to grade by the first method employed may be made from Schmitt's results. The author gives, in Table I, II, III, IV, V, VI and VII on pages 70, 71, 73, 74, 75, 76 and 77 of her monograph, the results of each subject in each grade on each test. From these tables the present writer has calculated the percentage passed in each test. A study of this sort rests for its reliability on the accuracy of the published tables, and the facts indicated by the tables do not always coincide

with Schmitt's discussion.[2] The writer has followed the tables rather than the discussion in calculating the results. In the VIII-2 test where an alternative rank is given for counting from 10 to 0 instead of 20 to 0, the writer has considered success in counting from 10 to 0 as a failure in counting from 20 to 0. In the line suggestion test Schmitt recognizes two types of failure, the typical failure according to Binet of accepting the suggestion of the first three lines, and the failure due to the fact that the subject actually judges the lines unequal after studying them. The second type of response Schmitt marks as passed, using a special symbol. The writer has calculated these percentages separately, entering the first or Binet type of response under "Line suggestion A" in the table, and the second type under "B." The V year and Adult tests were omitted. All of the other tests were included that had been given over 70% of the possible number of times. Unless otherwise noted, each test was given 100% of the possible number of times. Table 8 shows the per cent. that Schmitt's subjects in each grade passed each test in Binet's 1911 scale (Town's translation with modifications). The table is given with the reservation that the tables from which the percentages were calculated might contain misprints, and that the writer's interpretation of the tables might be at fault.

Inasmuch as there are many differences in procedure in giving the tests, and in the character of the schools tested, the results of the two investigations are not comparable in respect to the percentage passed in one grade in one study with those in the same grade in the other study. The method used in determining the

[2] In the discussion (page 69) Schmitt gives 15 subjects in the kindergarten failing test VII-4. Table I shows 13. On the same page she gives 24 subjects failing VIII-4. Table I shows 22 failing. In discussing the results of Grade I (page 72) Schmitt states that there is "more than 50% of failure with the discrimination of weight", while Table II shows 35% failure. Again, the tests referred to specific school instruction by Schmitt are VII-4, VIII-4, and IX 1, 2, 3 and 4. On page 72, in discussing the results of Grade I, she says "the tests below ten years which depend upon specific instruction are usually not passed except the VII-4 test. The percentages passed are as follows: VII-4 = 85%; VIII-4 = 45%; IX-1 = 35%; IX-2 = 75%; IX-3=90%; IX-4=30%. "Usually not passed" includes, therefore, tests passed 75% and 90% of the time.

TABLE 8

Per cent. that Schmitt's Subjects of Each Grade Passed Each Test. 150 Subjects.

	Grades						
	K	I	II	III	IV	V	VI
Number of subjects	25	20	17	21	22	22	23
VI-1, Distinguishing morning, afternoon	96	100*					
2. Defining in terms of use	92	94*					
3, Copying diamond	76	94*					
4, Counting 13 pennies	92	100*					
5, Choosing prettier of faces	92	100*					
VII-1, Showing right hand	92	80	100				
2, Describing pictures	72	65	81				
3, Executing 3 commissions	92	95	100				
4, Counting stamps	48	85	100				
5, Naming colors	96	100	100				
VIII-1, Comparing remembered objects	92	100	100	100	100		
2, Counting backwards from 20 to 0	40	85	94	95	100		
3, Indicating omissions in pictures	100	95	94	100	100		
4, Giving day and date	12	45	94	100	100		
5, Repeating 5 digits	64	85	94	100	100		
IX-1, Making change	6*	35	71	95	86	100	
2, Defining in terms superior to use	39*	75	65	100	95	100	
3, Naming pieces of money	28*	90	94	100	100	100	
4, Naming the months	6*	30	71	95	95	95	
5, Comprehending easy questions	61*	100	100	95	100	100	
X-1, Arranging 5 weights		65	41	57	50	64	
2, Copying designs		10	35	57	45	32	
3, Detecting absurdities		60	88	100	100	100	
4, Comprehending difficult questions		85	100	100	100	100	
5, Constructing sentence. Two ideas		65	76	100	100	100	
XII-1, Resisting suggestion, A. (Binet scoring)		64*	76	52	41	14*	100
B. Judgment error counted plus				100	86	100*	
2, Constructing sentence. One idea		57*	71	95	95	100*	100
3, Giving 60 words in three minutes		43*	82	62	100	95*	96
4, Defining abstract terms		7*	29	52	73	95*	100
5, Reconstructing dissected sentences		0*	6	10	23	81*	78
XV-1, Repeating 7 digits						62*	78
2, Rhyming words with "obey"						86*	70
3, Repeating a sentence of 26 syllables						10*	17
4, Interpreting pictures						14*	70
5, Solving problems from various facts						62*	70

Note.—All tests except those marked (*) were given all the possible number of times. The VI year tests were given 90% of the time in Grade I, the IX year tests 72% of the time in the kindergarten, the XII year tests 70% of the time in Grade I, and the XII and XV year tests 95% of the time in Grade V.

correlation of the tests with grade is the same as that used in the first method of treating the Princeton data, that of comparing the differences between grades by one-grade and two-grade steps, of selecting an arbitrary standard for detecting exceptional growth, and of comparing the resulting lists. The differences between the performance of each grade and the next succeeding grade were calculated. These differences, 100 in number, ranged from —24% to +62%, the median being +5% (Q=10.75%). The run of differences differs from that found in the Princeton study in two respects, in having a lower median and variability, and in containing more minus deviations. The lower median and variability is due to the fact that the tests were given over a wider range, the Princeton tests being given only on the "up slope" of the growth curve, or not being given when the tests were any distance above or below the probable range of ability of the group. The Princeton results showed only 4 minus deviations of 4, 3, 2, and 1% respectively, while Schmitt's results show 15 such deviations, 6 of them being 10% or over. These deviations are probably due to the smaller number of subjects, and if due to chance, should be counteracted by the precautionary measure of combining the indices of correlation into two-grade steps. 71 two-grade differences were obtained ranging from — 25% to + 82%, the median being + 10% (Q=16.5%). 4 measures were still in the minus direction, one of these, —25% (Design III to V) is probably significant, the other values of —6%, —5% and —4% having no significance. Inasmuch as the variability of the series is lower, those differences were considered to be worthy of special study that had the value of 2Q+M, or were in excess of the interquartile range plus the median. The lists of tests that appear as showing marked growth with grade according to the two methods are as follows:

One grade differences higher than 2Q+M	Two grade differences higher than 2Q+M
+62%, IX-3, Money, K to I	+82%, VIII-4, Date, K to II
+58%, XII-5, Dissected, IV to V	+71%, XII-5, Dissected, III to V
+49%, VIII-4, Date, I to II	+66%, IX-3, Money, K to II
+45%, VIII-2, 20 to 0, K to I	+65%, IX-4, Months, K to II
+41%, IX-4, Months, I to II	+65%, IX-4, Months, I to III
+39%, IX-5, Comprehension, K to I	+65%, IX-1, Change, K to II
+39%, XII-3, 60 words, I to II	+60%, IX-1, Change, I to III
+38%, XII-3, 60 words, III to IV	+55%, XII-5, Dissected, IV to VI
+37%, VII-4, Stamps, K to I	+55%, VIII-4, Date, I to III
+36%, IX-2, Definitions, K to I	+54%, VIII-2, 20 to 0, K to II
+36%, IX-1, Change, I to II	+52%, VII-4, Stamps, K to II
+35%, IX-2, Definitions, II to III	+47%, X-2, Design, I to III
+33%, VIII-4, Date, K to I	+45%, XII-4, Abstract Def., I to III
+29%, IX-1, Change, K-I	+44%, XII-4, Abstract Def., II to IV
+28%, X-3, Absurdities, I to II	+43%, XII-4, Abstract Def., III to V

A study of the above lists shows, as in the similar study of the Princeton data, that although the method of selecting the exceptional tests is an arbitrary one, the method is justified in practice, for only a few tests (13) appear in the lists as significant. In all, there were 34 tests[3] studied, and 30 differences were considered large enough to be significant. These 30 differences were confined to 13 tests. The tests of naming 60 words and defining in terms of use drop out of the first list owing to the elimination of the errors of negative correlation. The design test is both positive and negative, the ability increasing from Grades I to III and decreasing after III. The test of defining abstract terms appears according to the second method because the ability increases with grade from 7% in I to 95% in V by

[3] No differences were calculated from the line suggestion test owing to the possibility of misinterpreting the symbols. Schmitt notes the difference in the character of the responses from the suggestion error to the judgment error in passing from Grade II to III. The scoring of the suggestion error in the tables shows an inverse correlation with Grades II, III, IV and V, and a sudden change again from 14% in Grade V to 100% in Grade VI, so that there is probably a mistake. The scoring of the responses to this test according to the strict Binet ruling would make the "mental ages" lower, for many cases would then have basal X.

increases of approximately 25% in each grade. No conclusions may be drawn concerning the easy comprehension test and the absurdities test. The 20 remaining differences are confined to 7 tests, those of naming the day and date, naming the months, counting from 20 to 0, counting stamps, naming money, reconstructing dissected sentences, and making change. The first four were included in the five found to show the most marked influence of grade in the Princeton study. The test of naming the pieces of money did not show a marked relation to grade in the latter study, but this difference might be one of school curriculum. The test of naming the days of the week is not included in Binet's 1911 scale.

In the Princeton study alternatives were used in the making change question so that no data from this test were included in the quantitative study. These data show the ability in this test developing in the second and third grades, the test being passed only twice in the kindergarten and first grades, and generally passed above the third. The data in the test of reconstructing dissected sentences show very few passing the test below grade V with approximately three fourths passing in V and VI. In so far as the Trenton experimenting was applied to a few subjects in the regular grades below the seventh, this test was rarely passed in the third and fourth grade, passed about 5% in V, and almost universally passed in VI, VII and VIII. The number of subjects in each grade is small in the Trenton experiment, but each test was separately scored, i.e. each part of the dissected sentence test, each part of the absurdity test etc. Each of the three parts of the dissected sentence test showed the same growth between the same grades, and this growth was more marked than that in any other test. The evidence concerning these two tests, therefore, supports the evidence from Schmitt's results.

The quantitative analysis of the Princeton data and Schmitt's data would indicate that the tests of counting stamps, counting from 20 to 0, naming the days of the week, giving the day and date, naming the months, naming the pieces of money, making change and reconstructing dissected sentences were influenced to a considerable extent by grade training. The performance in

certain of these tests (days, date and months) may be the result of specific school training in the tests themselves, while others (perhaps the tests of counting stamps, counting from 20 to 0, and reconstructing dissected sentences) may involve a transfer effect in the application of the content of the grade in a new way. The fact that the tests correlate very highly with grade training does not show that the tests are worthless, but it does show that they should, perhaps, be placed in another scale, or should at least be placed on a different footing than those that test capacity irrespective of attainments.

One of the best tests[4] of intelligence is the determination of what an individual can do with the training he has received, but tests of this sort rest on the assumption that the individual's opportunities have been determined. The importance of tests of information in cases of alienation presenting a picture of deterioration is recognized. The important change to be made is not the elimination of such tests from intelligence scales, but their standardization on a different basis. The diagnostic value of such tests rests not in the mechanical memorizing of a time series such as that of the months, but in the ability to apply such a series. In pointing out this fact Katzenellenbogen (37) suggests that the months test be given in some such manner as "If somebody asks you in November to return three months later, what month would it be?" Decroly and Degand also suggest that the mechanical tests of counting and naming the days of the week and months be modified in some such manner.

[4] The writer recalls two cases in which the failure in tests which involved the application of training was very significant. The first was that of a woman of about 30, a parole patient in a hospital for the insane, who had never shown any marked symptoms other than a history of intellectual inferiority. This patient passed practically all of the Binet tests in the IX, X and XII year groups, but failed completely in the test of making change. This observation was later checked up. Another case of a woman of 22, in the same hospital, presented a border-line psychoneurotic picture perhaps, but no marked symptoms other than a history of intellectual inferiority. She passed in a great many of the difficult tests in the upper years but had great difficulty in telling time. Both cases had lived under very good home conditions and had mingled with people of ability. A great many tests of capacity were given, but the most illuminating evidence of their mental status came from the two tests mentioned.

Comparing the conclusions of this study with other investigations, the agreement is fairly close. Schmitt's results do not support her suggestion that the definitions test relates to specific school instruction. The other tests which she refers to this factor (stamps, date, 20 to 0, change, months and money) show the influence to a marked extent. Binet in classifying some of the tests referred the tests of copying a sentence, reading for memories, writing from dictation, copying a diamond, counting backwards and making change to scholastic training. The first three tests were not included in this investigation. The diamond test showed the influence of age to be as great if not greater than that of school training. The last two tests showed a marked influence of training. Binet referred the tests of counting 13 pennies, naming four colors, naming the days of the week and enumerating the months to home training. The last two showed a marked influence of school training. The results of the present investigation agree with those of Chotzen in finding no effect or very little effect of training in the tests of copying the diamond, repeating digits, describing pictures, counting 13 pennies, naming colors, comparing remembered objects, defining in terms of use and superior to use, and in finding marked influence of this factor in the test of naming the days of the week.

The methods used in analysing the results, especially the second method, reveal several suggestive relations between the tests and the school grades. There is a general correlation between the tests and the grades, a correlation that is very necessary to establish, for there is also a general correlation between intelligence and grade. In analysing the results of the individual tests by comparing the results of subjects of the same age in different grades, and of subjects of different ages in the same grade (Table 7), it was seen that, as a general rule, the growth in any particular ability occurred in passing from grade to grade, not in passing from age to age within one grade. In fact in only half of the cases in which the subjects of two ages in one grade may be compared do the older subjects make records that are higher than those of the younger ones, and only 10% of these gains are over 20%. If the groups were considered to be equal in all

cases in which their records were within 10% of each other, equality occurs in exactly 50% of the cases. Of the remainder, 20% of the groups were lower, while in only 30% of the cases are the older subjects actually higher than the younger subjects of the same grade. Some of the cases of retrogression could well be accidental, but they occur too frequently to be due entirely to chance.

Applying the same general method to the cases in which groups of the same age but in different grades were compared, 5% of the groups in a higher grade showed lower scores, the results correspond in 43% of the cases, while 52% showed definite improvement. This might indicate that there is a higher correlation between the tests and grade than between the tests and age. The fact that the comparison of children of different ages in the same grade showed the older children making lower records in 20% of the cases, equal records in 50% of the cases and higher records in only 30%, would confirm the general diagnostic value of the tests if Bonser's interpretation of this phenomena is correct. Bonser (12) applied various sorts of reasoning tests to children in the fourth, fifth and sixth school grades. In summarizing the results of the tests in the different grades, he says, "In the contrast with grade progress and progress with age, in the generally superior showing made by the younger groups of children of any grade when contrasted with the older pupils of the grade, and in the fairly substantial percentage of pupils from lower grades found in the highest quartile of ability for all, it is shown that native capacity is measured to a high degree by the tests."

In conclusion, the results shown in this chapter would indicate a correlation between the individual tests studied and the school grades, this correlation being high enough in some cases to show the actual effect of training. In answer to the general objection that since one demonstration of the accuracy of the tests rests on their correlation with school grades, the school grades are the real measure of intelligence and the mental tests superfluous, it is only necessary to point out that intelligence tests, besides affording the opportunity for accurate standardization,

also detect the subject's potential abilities independent of his past performance. The school measure indicates mental defect in cases of gross retardation, but it does not indicate exceptional ability.

Schmitt's contention that the school represents a standard environmental situation, and a measure of a subject's ability should include a measure of the adequacy of his reaction to this situation, is well founded. It is not, however, a criticism of the Binet scale, for the scale aims to test native capacity. At the Buffalo conference (15) on the Binet scale, the following question was raised,—"What is it, after all, that the scale aims to test?" The question was answered by "We believe that current misconceptions as to the aim of the scale should be removed. It is not intended to test the emotional or volitional nature, but primarily intelligence (judgment)." To this list might be added the assertion that the scale was not intended to test a child's reaction to the school situation, or to furnish an outline for taking a record of his life history.

Rogers and McIntyre (54) would also have mental tests include tests dependent on both school and home training. This general trend of present day discussion is a reversion to Binet's 1908 type of scale, a tendency to which Binet was in opposition. The probable solution rests in eliminating from the scale the tests involving training, and in constructing a standardized scale of another sort for the estimation of the individual's reaction to the school situation in terms of the length of time that he has met that situation. That such a scale is not a matter of speculation is shown by the number of scales now on the market for measuring handwriting, spelling, composition, arithmetical ability, etc. Tests of native capacity and tests dependent on school and environmental training cannot be standardized on the same basis, for they are essentially different measures. Measures of the first sort may perhaps be correlated with age, while measures of the other sort can be correlated only with opportunity.

V. SEX DIFFERENCES

The investigators who have studied the influence of sex differences on the Binet-Simon tests have used two methods, that of comparing the "mental ages" or total scores of subjects of each sex, and that of comparing the per cent. that the subjects of each sex pass each test. The first method throws no light on the individual tests, inasmuch as one sex may be superior in one test and inferior in another so that the total score will balance the influence of this factor. Inasmuch as the scale is founded on the principle that sex differences do not exist, it is important to study the individual tests, and to determine the accuracy of this assumption.

The Princeton data are available for a study of this sort. 352 subjects (187 boys and 165 girls) between the ages 6 and 12 were examined. The method of study adopted was that of comparing the results of non-selected boys and girls of each age, and, as a check on this method, of comparing the results of selected boys and girls of four ages.

Inasmuch as the subjects of each chronological age are distributed over a range of one year (the 6 year subjects for example being distributed from 6.0 to 6.9), the actual average age of the subjects of each age was computed to make sure that no differences might appear due to the chance selection of subjects at either extreme. These averages are shown in Table 9.

TABLE 9

Actual Average Chronological Age of Boys and Girls in Each Age Group.

	BOYS		GIRLS	
	Number of Subjects	Average Age (M. V.)	Number of Subjects	Average Age (M. V.)
Age 6	37	6.58 (0.20)	23	6.51 (0.20)
Age 7	29	7.50 (0.29)	31	7.39 (0.26)
Age 8	24	8.48 (0.29)	28	8.48 (0.22)
Age 9	20	9.46 (0.27)	22	9.54 (0.26)
Age 10	31	10.46 (0.25)	23	10.37 (0.30)
Age 11	28	11.59 (0.22)	20	11.52 (0.27)
Age 12	18	12.43 (0.30)	18	12.57 (0.24)

A perusal of this table shows that the subjects agree closely both in their average and in their variability. The 12 year boys are actually 0.14 yr. younger than the girls of the same age group. The 7 year boys are 0.11 yr. older than the 7 year girls. All other differences are less than 0.10 yr. The correspondence is close enough for all practical purposes, but these differences must be taken into consideration before drawing final conclusions.

The 352 non-selected subjects from 6 to 12 were distributed throughout the kindergarten, special class, and first six minus and plus grades as shown in Table 10.

TABLE 10

Age in Grade Distribution of 187 Boys and 165 Girls, 6 to 12 Years of Age.

Age	6		7		8		9		10		11		12		Totals
Sex	B	G	B	G	B	G	B	G	B	G	B	G	B	G	
Special Class	2		1		3		3				3		1		13
Kindergarten	8	3													11
Grade I-	13	4	8	9	1	1	1								37
Grade I	14	14	9	9	3	1			1						51
Grade II-			2	2	2	4		2							12
Grade II		2	7	10	6	9	2	3			1				40
Grade III-						3	4	2	1	1		1			12
Grade III			2		8	10	5	8	5	5	1	1			45
Grade IV-								1	5	1	2	2	3	1	15
Grade IV				1	1		5	6	10	4	1	2	3	2	35
Grade V-									1	1		1	3	3	9
Grade V									7	11	10	6	2	6	42
Grade VI-												1		1	2
Grade VI									1		10	6	6	5	28
Totals	37	23	29	31	24	28	20	22	31	23	28	20	18	18	352

It is generally conceded that a difference exists in the reactions of the sexes to the school curriulum, the girls in the long run making better progress in school work than the boys. A study of Table 10 shows that in general the girls have a slightly higher distribution than the boys, these relations being more clearly indicated in Table 11 in which the average grade of the subjects of each age and sex is shown. In computing the average grade, the kindergarten was counted 0; Grade I—, 0.5; Grade I+, 1.0; Grade II—, 1.5; etc. Each subject in the special class was assigned a grade 0.5 lower than the lowest subject of his age (0 being the smallest value given), on the theory that each subject in

the special class was less satisfactory than any of his comrades in the regular class. The fact that there were no girls in the special class would cause an unduly exaggerated difference between the average grades of the boys and girls. For this reason, the average grades of the boys, including and excluding the special class cases, were separately figured, these values being separately shown in Table 11 under Boys A (the average grade including the special class cases), and Boys B (the average grade excluding the classes). Had the special class subjects been in the regular grades, they would have lowered the average of each group, so that the two values may be taken only as limits, the values under "Boys A" being the lower limit, and those under "Boys B," the upper limit.

TABLE 11

Actual Average Grade of Boys and Girls in Each Age Group.

	BOYS A		BOYS B		GIRLS	
	No.	Average Age (M. V.)	No.	Average Age (M. V.)	No.	Average Age (M. V.)
Age 6	37	0.55 (0.34)	35	0.59 (0.33)	23	0.87 (0.35)
Age 7	29	1.24 (0.64)	28	1.29 (0.63)	31	1.31 (0.65)
Age 8	24	1.94 (0.91)	21	2.21 (0.65)	28	2.25 (0.59)
Age 9	20	2.48 (1.04)	17	2.91 (0.69)	22	2.98 (0.62)
Age 10	31	3.92 (0.71)	31	3.92 (0.71)	23	4.19 (0.80)
Age 11	28	4.66 (1.20)	25	5.04 (0.77)	20	4.83 (0.88)
Age 12	18	4.72 (0.91)	17	4.82 (0.88)	18	5.03 (0.59)

Table 11 shows that the scholastic ability of the girls as indicated by the average grade is uniformly higher than that indicated by the lower limit of the boys, and is below the upper limit of the boys in only one case (at 11 years). A slight sex difference in school work in favor of the girls may therefore be assumed at the outset. It is significant that the upper limit of the 11 year boys is higher than that of the 12 year boys, and that the lower limits show a difference of but 0.06. This would indicate a poor selection of 12 year boys, or a superior selection of 11 year boys. Both measures of the scholastic ability of the boys show a generally higher variability than that of the girls.

From Table 9 it may be seen that the growth in the actual average age of each sex is not uniform from year to year, the minimum increase for boys being 0.84 yr. (from 11 to 12), and

for girls 0.83 yr. (from 9 to 10), while the maximum increase for boys is 1.13 yr. (from 10 to 11), and for girls 1.15 yr. (from 10 to 11). A more marked lack of regularity in the growth of scholastic ability from year to year as measured by the average grade is shown in Table 11, no increase being shown by the 12 year boys over the 11 year boys, while the 10 year boys show an increase of 1.44 to 1.01 grades over the 9 year boys. In the same way the 10 year girls show an increase over the 9 year girls that is nearly three times that of the 7 year girls over the 6 year girls, while the increase of the 7 year girls over the 6 year girls is twice that of the 12 year girls over the 11 year girls. These relations indicate that the selection of subjects is not uniform at each age. The subjects of any one age may be either a superior or inferior selection of all children of that age, and there is no reason for supposing that this random sample of superior or inferior subjects of any age will correspond to a similar sampling of the subjects of the opposite sex of the same age.

The process of calculating the percentage that the boys and girls of each age pass each test is extremely simple, but the conclusion, that the differences found between the percentage passed by the sexes at each age may be attributed to sex differences, is not justified unless all the variable factors are known.

A previous chapter showed variations in the tests due to the influence of the personal equation of the experimenters. To avoid this variable influence, only those tests were studied that showed that they were free from the influence of this factor. Inasmuch as each experimenter examined approximately the same number of boys and girls of each age, any influence of this factor would be equalized, provided, of course, that there were no differences in the reaction of the experimenters to the two sexes. In the detailed study of the design test, it was found that experimenter C was more lenient in marking girls than boys. The possibility of a similar interpretation in a few other tests was suggested, but not demonstrated. In analysing the results for sex differences, however, the possibility of such an interpretation must be kept in mind.

Another possible source of error is that due to incomplete data.

The experimenters, in giving the tests, would give only those within the approximate range of the subject, so that each test would be given to a superior selection of children below the normal range of the test, and to an inferior selection of subjects above this range, a process tending to make the apparent growth of an ability less than the probable real growth. In comparing the results of the sexes, however, it is not necessary to have accurate results on the growth of an ability, but results which have the same determining factors. If the experimenters gave the test to approximately the same proportions of boys and girls at each age, a comparison of the percentage passed is legitimate, even if a small proportion of the whole group were actually tested, for the proportion would include the same selection of subjects. The number of boys and girls at each age, and the percentage that each test was given to these subjects are shown in Table 12. The test of counting 13 pennies, for example, was given 37 times to 6 year boys, or 100% of the possible number of times, while the test of counting from 20 to 0 was given 27 times to the same group, or 73% of the possible number of times. Column A shows the total number of times each test was given to all of the boys and girls. Column B gives the average age of all the boys and girls to whom each test was given. The average given in this case is not the actual average derived from the actual chronological age of each subject figured in tenths, but the weighted[1] average, the whole numbers 6, 7, 8, 9, 10, 11, and 12 being used.

Table 12 shows a very close correspondence between the percentage that each test was given to boys and girls of each age, so that the error due to incomplete data, though present, is present to the same extent in the results of both sexes, and may be disregarded. A fairly close correspondence in the average age of all the boys and girls to whom each test was given is also indicated in Table 12. In the test of counting stamps there is an

[1] For example, in the test of counting 13 pennies, the average age of the boys to whom the test was given is,—

$$\frac{(37\times6)+(28\times7)+(16\times8)+(8\times9)+(7\times10)+(3\times11)+(1\times12)}{100} = 7.33 \text{ years}$$

TABLE 12

Percentage that Each Test Was Given to Boys and Girls of Each Age, the Total Number of Times Each Test Was Given to Each Sex and the Average Age of All Subjects of Each Sex to Whom Each Test Was Given.

Chronological age		6	7	8	9	10	11	12	A Total number of times given	B Average age of subjects. (weighted)
Number of subjects	Boys	37	29	24	20	31	28	18		
Number of subjects	Girls	*23*	*31*	*28*	*22*	*23*	*20*	*18*		
Counting 13 pennies	Boys	100	97	67	40	23	11	6	100	7.33
	Girls	*100*	*94*	*68*	*41*	*26*	*10*	*11*	*90*	*7.56*
Describing pictures	Boys	100	90	67	45	26	11	6	100	7.38
	Girls	*100*	*94*	*68*	*41*	*30*	*20*	*11*	*93*	*7.66*
Copying diamond	Boys	100	93	63	60	32	14	17	108	7.30
	Girls	*100*	*94*	*61*	*64*	*35*	*20*	*11*	*97*	*7.74*
Naming colors	Boys	100	93	67	45	23	11	6	100	7.35
	Girls	*100*	*94*	*68*	*41*	*30*	*15*	*11*	*92*	*7.62*
Counting from 20 to 0	Boys	73	97	83	80	61	21	44	124	8.18
	Girls	*74*	*71*	*79*	*77*	*52*	*35*	*28*	*102*	*8.25*
Counting stamps	Boys	65	97	88	80	61	21	44	122	8.23
	Girls	*83*	*87*	*79*	*82*	*52*	*35*	*39*	*112*	*8.23*
Repeating all digits	Boys	95	100	100	100	100	100	100	185	8.75
	Girls	*96*	*97*	*96*	*100*	*100*	*100*	*100*	*162*	*8.78*
Naming days of week	Boys	92	100	83	80	61	21	44	132	8.04
	Girls	*96*	*90*	*82*	*82*	*52*	*35*	*28*	*115*	*8.10*
Giving day and date	Boys	43	76	88	95	84	64	89	138	8.10
	Girls	*78*	*77*	*93*	*100*	*78*	*75*	*72*	*136*	*8.70*
Naming the months	Boys	41	79	79	95	81	54	78	130	8.90
	Girls	*39*	*65*	*93*	*100*	*70*	*65*	*61*	*117*	*8.84*
Naming money	Boys	27	62	67	90	97	64	78	124	9.21
	Girls	*43*	*39*	*86*	*95*	*100*	*80*	*67*	*118*	*9.11*
Copying designs	Boys	16	31	67	85	94	79	78	113	9.56
	Girls	*26*	*19*	*57*	*86*	*96*	*80*	*67*	*97*	*9.46*
3 words in sentence	Boys	8	31	63	90	100	93	100	120	9.79
	Girls	*26*	*26*	*68*	*86*	*100*	*95*	*94*	*111*	*9.36*
60 words in 3 minutes	Boys	11	21	38	70	81	93	89	100	9.92
	Girls	*30*	*10*	*32*	*50*	*74*	*90*	*78*	*79*	*9.75*
Giving rhymes	Boys	8	21	25	50	74	89	94	90	10.08
	Girls	*13*	*13*	*36*	*45*	*74*	*90*	*83*	*77*	*9.92*
Defining "fork" etc.	Boys	38	62	50	55	48	25	17	80	8.35
	Girls	*61*	*65*	*61*	*68*	*39*	*20*	*11*	*81*	*8.09*

actual correspondence. The greatest difference is that of 0.6 yr. in the test of giving the date. The differences, on the whole, are small, but must be taken into consideration when comparing the percentages that all boys and girls pass each test.

Two methods are available for studying the influence of sex differences on the individual tests. The first is that of comparing the results of boys and girls of each age on each test. This method is affected by the chance selection of superior or inferior subjects, and the results can have no meaning unless the relations of the groups of each age of the same sex are understood. For example, the fact that the 12 year boys are 36% lower than the 12 year girls in the test of naming the months has no significance as an isolated finding, for its significance is modified by the additional fact that this group of 12 year boys is 10% lower than the 9 year boys, 12% lower than the 10 year boys, and 9% lower than the 11 year boys on the same test.

The other method is that of comparing the per cent. that all subjects of each sex pass each test. This method avoids the factor of variations in the results due to a chance superiority of one age group over the other of the opposite sex, but, at the same time, it tends to obscure the magnitude of the differences that might occur. The most reliable differential measure between two groups is one that is well within the range of ability of the groups. The difference will be obscured if the measure is too easy or too difficult. A comparison of the results of all subjects would, in this way, tend to minimize[2] the magnitude of the real difference between the groups. Furthermore, there is a possibility that one sex might acquire an ability first, but eventually be surpassed by the other. The per cent. that all subjects passed would show no deviation, because the two tendencies would balance.

[2] For example, if there were 20 subjects of each age and of each sex from 6 to 12, and a certain test were passed by 75% of the 6 year girls, and by all of the 7, 8, 9, 10, 11 and 12 year girls, by 50% of the 6 year boys, 75% of the 7 year boys and all of the remaining groups, the total percentage passed for all girls would be 96%, and for all boys, 89%. The differential character of the test is indicated by the value 7%, while its actual differential character, just within the range of ability of the groups, is 25%.

Neither method, then, is entirely satisfactory, the first because it would tend to exaggerate chance differences, the second because it would tend to obscure real differences. The method used in this study is that of comparing the results of non-selected and selected subjects of each age and sex, studying first the general growth of each ability from age to age within each sex, and using the per cent. that all subjects pass each test to determine the correlation between the results of non-selected and selected subjects.

Table 13 shows the percentage of proportion[3] that the boys and girls of each age pass each test, the percentage that all boys and girls pass each test, the actual percentage that the boys are superior to (+) or inferior to (—) the girls of each age, the difference between the average age of all boys and girls to whom each test was given, and the difference between the percentage that all boys and girls pass each test.

The differences between the performance of the boys and girls at each age have no meaning unless the general growth of the abilities in each sex is first understood. Studying first the results of the 187 non-selected boys shown in the first seven columns of Table 13, it may be seen that the growth of ability in each test is rather irregular. The test of naming the months, for example, shows a slight decrease from 9 to 12. The differences between the percentage performances of the subjects of each age and those of the preceding age were calculated. The 12 year group, compared to the 11 year group, is +11% on the test of giving the date, —9% on the test of naming the months etc. 61 differences were thus obtained, varying in magnitude from —15% to +36%, the median being +8% (Q=9.75%). 13 of the deviations (21%) were minus values. The largest negative deviations occurred in the tests of naming colors (—15%, 7 to 8), naming money (—15%, 11 to 12), and constructing a sentence containing two ideas (—13%, 8 to 9). The remaining 10 minus deviations were less than 10%.

[3] The proportion given is the number of times a test was given over the number of times a test was passed. No percentages were calculated for tests given less than 12 times, and no percentages are given for the definitions tests on account of the small number of times they are given to all subjects.

TABLE 13

Percentage or Proportion that the Boys and Girls of Each Age Passed Each Test, and the Differences between the Percentages Passed.

		Chronological Ages							% passed all subjects	Differences in percentage passed at each age							Difference in average ages	Difference in % passed by all
		6	7	8	9	10	11	12		6	7	8	9	10	11	12		
Counting 13 pennies	Boys	84	93	94	8/7	7/7	3/3	1/1	90	−16	− 4	− 6					−.23	− 9
	Girls	*100*	*97*	*100*	*9/9*	*6/6*	*2/2*	*2/2*	*99*									
Describing pictures	Boys	86	100	94	9/9	8/8	3/2	1/1	93	− 1	+17	+10					−.28	+ 5
	Girls	*87*	*83*	*84*	*9/9*	*7/7*	*4/4*	*2/2*	*88*									
Copying a diamond	Boys	57	93	93	92	10/9	4/2	3/3	79	−21	+17	+22	+13				−.44	+ 1
	Girls	*78*	*76*	*71*	*79*	*8/7*	*4/4*	*2/2*	*78*									
Naming four colors	Boys	70	96	81	9/9	7/7	3/1	1/1	83	−30	+ 3	−19					−.27	−14
	Girls	*100*	*93*	*100*	*9/8*	*7/7*	*3/3*	*2/2*	*97*									
Counting from 20 to 0	Boys	4	25	55	50	79	6/3	8/7	42	−25	− 7	+14	− 9	+ 4			−.07	− 8
	Girls	*29*	*32*	*41*	*59*	*75*	*7/6*	*5/5*	*50*									
Counting stamps	Boys	25	29	62	75	79	6/3	8/7	52	+14	+ 7	+ 7	− 3	+21			0	+ 6
	Girls	*11*	*22*	*55*	*78*	*58*	*7/5*	*7/6*	*46*									
Repeating all digits	Boys	12	28	35	56	59	63	73	43	− 9	− 7	− 9	+ 9	− 6	− 7	+ 1	−.03	− 6
	Girls	*21*	*35*	*44*	*47*	*65*	*70*	*72*	*49*									
Naming days of week	Boys	35	59	75	88	95	6/5	8/8	67	−24	−16	− 8	−12	− 5			−.06	−16
	Girls	*59*	*75*	*83*	*100*	*100*	*7/7*	*5/5*	*83*									
Giving day and date	Boys	13	18	48	84	92	89	100	64	− 4	− 7	− 6	− 7	− 8	−11	0	−.60	− 1
	Girls	*17*	*25*	*54*	*91*	*100*	*100*	*100*	*65*									
Naming the months	Boys	7	17	42	74	76	73	64	51		+ 2	−23	+ 6	−18	−19	−36	+.06	−13
	Girls	*9/2*	*15*	*65*	*68*	*94*	*92*	*100*	*64*									
Naming pieces of money	Boys	10/2	11	38	33	60	72	57	44		+11	+17	0	− 1	+ 3	−18	+.10	+ 4
	Girls	*10/1*	*0*	*21*	*33*	*61*	*69*	*75*	*40*									
Copying designs from memory	Boys	6/0	9/3	25	35	24	36	29	28			0	+24	−12	− 8	−21	+.10	0
	Girls	*6/0*	*6/0*	*25*	*11*	*36*	*44*	*50*	*28*									
3 words in sent. Two ideas	Boys	3/1	9/3	80	67	74	73	78	70			+17	+ 4	+ 9	−11	−10	+.43	+ 3
	Girls	*6/0*	*8/4*	*63*	*63*	*65*	*84*	*88*	*67*									
3 words in sent. One idea	Boys	*3/1*	9/3	27	22	45	58	67	44			+ 6	+ 1	+15	+21	+26	+.43	+18
	Girls	*6/0*	*8/0*	*21*	*21*	*30*	*37*	*41*	*26*									
Naming 60 words in 3 minutes	Boys	4/0	6/3	9/4	36	60	77	69	58				−19	+ 7	− 6	+19	+.17	+ 4
	Girls	*7/0*	*3/0*	*9/6*	*55*	*53*	*83*	*50*	*54*									
Giving rhymes with 3 words	Boys	3/0	6/4	6/5	10/4	57	60	65	58					− 2	− 1	−15	+.16	−10
	Girls	*3/1*	*4/3*	*10/8*	*10/7*	*59*	*61*	*80*	*68*									
Defining by use	Boys	14/13	18/16	12/12	11/10	15/15	7/6	3/3	94								+.26	+ 4
	Girls	*14/13*	*20/15*	*17/16*	*15/14*	*9/9*	*4/4*	*2/2*	*90*									
Defining superior to use	Boys	14/1	18/6	12/7	11/6	15/3	7/2	3/1	33								+.26	+ 7
	Girls	*14/3*	*20/4*	*17/5*	*15/5*	*9/3*	*4/1*	*2/0*	*26*									

An index of the growth from year to year was obtained by calculating the average percentage increase from one age group to another. For example, the 7 year boys were 26% higher than the 6 year boys in the test of naming colors, 5% higher in naming the date etc. The average of the 10 possible comparisons between 6 and 7 year boys shows that the latter averaged 16.1% higher than the former. The average increases in percentage passed from year to year are as follows,—6 to 7=16.1%; 7 to 8=13.5% 8 to 9=8.7%; 9 to 10=11.2%; 10 to 11=6.0%; and 11 to 12=0.2%. These figures show strikingly the irregularity of the growth from age to age. Comparing these average percentage increases in tests with the averages shown in Tables 9 and 11, there is no observable relation between this increase and the increase in average age from age to age, or the increase in average grade from age to age. The smallest increase in the tests (0.2%, 11 to 12) coincides with the smallest increase in average age from year to year (0.84 yr.), and the smallest increase in average grade from year to year. The other relations are varied.

The fact of the variability in the results of the non-selected boys stands out. The irregularity of the growth of the various abilities, and the fact that in 21% of the cases the boys of one age are actually lower than those of the previous age, point to the conclusion that certain allowances will have to be made for chance variations. It is not possible to acccount for the variations in growth by reference to the relative increase in average age or average grade from year to year.

The results of the 165 non-selected girls, shown in italics in the first seven columns of table 13, were studied in the same manner as the results of the boys. 60 differences between the percentage performance of the girls of each age and those of the preceding age were obtained. These differences ranged from —33% to +50%, the median being 7% (Q=8%). 10 of the deviations (17%), were minus values. The largest deviations were shown in the tests of naming 60 words, (—33%, 11 to 12), counting stamps (—20%, 9 to 10), and drawing designs

(—14%, 8 to 9). The remaining 7 minus deviations were below 10%.

The average increases in the percentage passed from year to year are as follows,— 6 to 7=3.9%; 7 to 8=15%; 8 to 9=8.8%; 9 to 10=10.1%; 10 to 11=8.7%; 11 to 12=1.8%. Both boys and girls show the smallest average increase in the percentage passed in the step from 11 to 12, and the magnitudes of the increases agree fairly well except for the step from 6 to 7. The increase of the 7 year girls over the 6 year girls is 3.9%, the next to the smallest increase of one age group over any preceding group. The 7 year boys, however, show an average increase of 16.1%, over the 6 year boys, the largest increase of any group of boys over any preceding group. It will be difficult, then, to draw conclusions concerning sex differences from a comparison of the 6 year boys and girls, for the 6 year girls are either a superior selection or the 6 year boys are an inferior selection if the character of these groups be judged by the comparison with the 7 year subjects. The same comparison, on the other hand, might indicate that the 7 year girls were an inferior selection and the 7 year boys a superior selection from the general run. It is only possible to point out the irregularity, however, it is not possible to show the cause of the irregularity.

A comparison of the average increase in the percentage passed by girls from age to age with the increase in the average ages shown in Table 9 shows no demonstrable relation to exist. Comparing this growth in the ability on the tests with the growth in average grade, shown in Table 11, shows a very positive relation to exist between these factors. Where the increase in average grade is smallest (i.e. from 6 to 7 and from 11 to 12), the increase in the tests is smallest (3.9% and 1.8%), while the greatest increase in grade (from 9 to 10 and from 7 to 8) coincide with the greatest increase in the test abilities (10.1% and 15.0%). This relation was not indicated in the results of the boys. The explanation of this fact that a correlation between the increase in the tests with grade was found in the results of the girls but not of the boys is a matter of speculation. It has been shown that the boys have a higher variability in grade than

girls. This tendency of the boys to be distributed in a wider range of grades might nullify the grade correlation slightly, but probably not to any considerable extent. The fact that the causes of this variation are not determined serves to illustrate the dangers of comparing the results of two groups when the factors operating on the groups are not known.

The foregoing study of the growth of the various abilities from age to age in each sex, and the analysis of the causes influencing this growth, demonstrates the great variability of the results. This fact of variability must be considered before drawing conclusions concerning sex differences by the method of comparing the results of boys and girls of each age.

The percentage differences between the performance of non-selected boys and girls of each age are shown in Table 13. In actual magnitude, these differences vary from 0% to 36%, the median being 9% (Q=5.5%). 75% of the differences are 17% or under, and only 16% are over 20%. In regard to sign, the differences vary from —36% to +26%, the median being —3.5% (Q=8.75%), showing a slight general superiority of the girls. If the number of possibilities of variation in comparing the results of small groups of non-selected subjects are taken into consideration, the presence of mental defectives, of subjects having language difficulties, of subjects in different grades influenced by different training, the possibility of a superior selection of subjects at one age group than at another, and the probability that similar chance samplings would not fall at the same age, the fact of correspondence indicated in Table 13 has more meaning than the fact of divergence.

The variability indicated in the study of the growth of abilities with age was so great that it makes interpretation of the results in terms of sex differences very difficult, and warranted conclusions impossible. It is legitimate to expect that the older subjects of either sex should make higher scores than the younger subjects of the same sex, but this was not found to be the universal rule. The boys' results showed minus deviations in 21% of the cases and the girls' results showed minus deviations in 17% of the cases. In one case the 12 year girls were 33% lower than

the 11 year girls. If this value (33%) be taken as the error due to chance variation, then only one value, that of —36%, (naming the months, age 12), may be taken as significant, and it has been seen that in this test the 12 year boys are 10% lower than the 9 year boys. The conclusion would follow, then, that there were no sex differences. This alternative, however, seems to place too much weight on one variation so that the truth probably lies in the assertion that the sex differences, that actually exist, are slight.

A study of the reactions of selected groups of boys and girls should throw light on the results from non-selected subjects, and make conclusions more certain. Subjects were selected by a process of elimination and selection. All of the subjects that were in the special class and minus grades were eliminated, along with all children of non-English speaking parents. From the following group of English speaking subjects in the regular grades all subjects were eliminated who had entered grade at an age very much above or below that of the general run of entrants.[4] The remaining subjects ranged in age from 4.3 years to 14.4 years, but were found to group rather closely around certain ages. It was possible to find four groups of boys and girls of approximately the same chronological ages. The character of these subjects is indicated in Table 14.

The four groups of subjects, chronologically from 6.0 to 6.9, 7.6 to 8.9, 9.7 to 10.9 and 11.7 to 13.3 (which will be referred to as 6, 8, 10 and 12), were distributed in approximately the same grades, and had approximately the same average age and average grade. The results of these groups are shown in Table 15, which is arranged to show all the facts for selected subjects that were given for non-selected subjects in Tables 12 and 13. The first four columns show the percentage that each test was given to each group. The next four columns show the percentage or the proportion that the subjects in each group passed each

[4] The ages on entering each grade of the subjects retained were as follows,—Kindergarten = 4, 5 and 6; Grade I = 5, 6 and 7; Grade II = 6, 7 and 8; Grade III = 8, 9 and 10; Grade IV = 9, 10 and 11; Grade V = 10, 11 and 12; Grade VI = 11, 12 and 13.

TABLE 14

Age in Grade Distribution, Average Grade and Average Age of 167 Selected Subjects. 86 Boys and 81 Girls.

Age Group	Sex	Age in Grade Distribution K	I	II	III	IV	V	VI	TOTAL	Average Grade (M.V.)	Average Age (M.V.)
6.0 to 6.9	Boys	5	13						18	0.72 (0.40)	6.52 (0.22)
	Girls	3	13	2					18	0.89 (0.39)	6.53 (0.22)
7.6 to 8.9	Boys		7	13	3				23	1.83 (0.51)	8.09 (0.38)
	Girls		2	13	5				20	2.15 (0.43)	8.32 (0.38)
9.7 to 10.9	Boys				6	12	2		20	3.80 (0.48)	10.37 (0.36)
	Girls				9	7	5		21	3.81 (0.69)	10.14 (0.32)
11.7 to 13.3	Boys					2	8	15	25	5.52 (0.58)	12.35 (0.55)
	Girls					3	8	11	22	5.36 (0.64)	12.41 (0.46)

test. Column A shows the total number of times each test was given to all boys and girls, Column B, the weighted average age (the average ages given in Table 14 being used), and Column C the percentage that all subjects passed each test. The next four columns show the percentage that the boys are above (+) or below (—) the girls. Column D (derived from Column B), gives the difference between the average ages of all subjects to whom each test was given. Column E (derived from Column C), gives the differences between the percentages passed by all boys and girls on each test.

The growth of the various abilities with age in the selected groups of subjects is more uniform than that shown by the non-selected subjects. Only three cases appear in which the younger subjects make higher scores than those of older subjects, these exceptions occurring in the tests of describing pictures (—3%, girls 6 to 8), naming colors (—7%, girls 6 to 8), and naming months (—9%, boys, 10 to 12). In the comparison of the sexes 41 differences are obtained varying in magnitude from —28% to +26%, the median being 0% (Q=9.5%). In actual magnitude the differences vary from 0 to 28, the median being 10% (Q=4.75%), the median being 1% higher than that of non-selected data, and the variability 0.75% less. 75% of the differences were less than 14%.

TABLE 15

Results of 167 Selected Subjects. (86 Boys and 81 Girls).

		Percentage test was given.				Percentage or proportion passed				Columns			Percentage that boys are higher or lower than girls				Columns	
		6	8	10	12	6	8	10	12	A	B	C	6	8	10	12	D	E
Counting 13 pennies.	Boys	100	87	15	0	94	95	3/3		41	7.57	95	— 6	— 5			—.25	— 5
	Girls	*100*	*70*	*24*	*5*	*100*	*100*	*5/5*	*1/1*	*38*	*7.82*	*100*						
Describing pictures.	Boys	100	83	20	0	100	100	4/4		41	7.38	100	+11	+14			—.44	+10
	Girls	*100*	*70*	*24*	*5*	*89*	*86*	*5/5*	*1/1*	*38*	*7.82*	*90*						
Copying a diamond.	Boys	100	83	40	0	56	100	8/8		45	7.87	82	—27	+ 8			—.16	— 8
	Girls	*100*	*60*	*43*	*5*	*83*	*92*	*9/9*	*1/1*	*40*	*8.03*	*90*						
Naming four colors.	Boys	100	83	15	0	89	95	3/3		40	7.55	93	—11	+ 2			—.27	— 4
	Girls	*100*	*70*	*24*	*5*	*100*	*93*	*5/5*	*1/1*	*38*	*7.82*	*97*						
Counting from 20 to 0.	Boys	94	96	75	12	6	36	93	3/3	57	8.45	46	—27	—28	+ 2		+.10	—16
	Girls	*83*	*70*	*52*	*9*	*33*	*64*	*91*	*2/2*	*42*	*8.35*	*62*						
Counting stamps.	Boys	78	100	75	12	14	48	93	3/3	55	8.50	55	+ 1	+ 8	+26		—.08	+11
	Girls	*89*	*75*	*57*	*23*	*13*	*40*	*67*	*5/5*	*48*	*8.58*	*44*						
Repeating all digits.	Boys	100	100	100	100	16	39	56	75	86	9.53	48	—11	0	—11	+ 4	+.01	— 4
	Girls	*100*	*95*	*100*	*100*	*27*	*39*	*67*	*71*	*80*	*9.52*	*52*						
Naming days of week.	Boys	100	100	75	12	44	78	100	3/3	59	8.41	75	—23	+ 9	0		+.14	— 2
	Girls	*100*	*80*	*57*	*9*	*67*	*69*	*100*	*2/2*	*48*	*8.27*	*77*						
Giving day and date.	Boys	67	96	90	60	8	27	100	100	67	9.37	60	—12	—12	0	0	+.03	— 5
	Girls	*83*	*90*	*86*	*68*	*20*	*39*	*100*	*100*	*66*	*9.34*	*65*						
Naming the months.	Boys	56	91	85	52	10/1	29	94	85	61	9.38	56		—12	+ 6	—15	—.20	—12
	Girls	*39*	*85*	*76*	*55*	*7/2*	*41*	*88*	***100***	*52*	*9.58*	*67*						
Naming pieces of money.	Boys	39	74	100	72	7/2	24	60	78	62	9.88	52		— 7	+ 8	— 4	+.03	+ 2
	Girls	*50*	*65*	*100*	*77*	*9/1*	*31*	*52*	*82*	*60*	*9.85*	*50*						
Copying designs from memory.	Boys	11	61	100	68	2/0	14/3	30	53	53	10.26	34			+ 6	—10	+.08	— 1
	Girls	*28*	*45*	*100*	*73*	*5/0*	*9/3*	*24*	*63*	*51*	*10.18*	*35*						
3 words in sent. two ideas.	Boys	6	57	100	100	1/1	13/6	75	88	59	10.64	75			— 4	—12	+.40	— 2
	Girls	*33*	*50*	*100*	*91*	*6/0*	*10/9*	*71*	*100*	*57*	*10.24*	*77*						
3 words in sent. one idea.	Boys	6	57	100	100	1/1	13/2	50	68	59	10.64	51			+17	+23	+.40	+18
	Girls	*33*	*50*	*100*	*91*	*6/0*	*10/3*	*33*	*45*	*57*	*10.24*	*33*						
Naming 60 words in 3 minutes.	Boys	11	30	80	96	2/0	7/3	69	88	49	10.85	71			+13	+20	+.52	+18
	Girls	*39*	*25*	*76*	*86*	*7/0*	*5/3*	*56*	*68*	*47*	*10.33*	*53*						
Giving rhymes with 3 words.	Boys	6	22	65	96	1/0	5/3	69	71	43	11.12	67			+ 6	— 3	+.49	+ 1
	Girls	*17*	*30*	*76*	*86*	*3/1*	*6/4*	*63*	*74*	*44*	*10.63*	*66*						
Defining by use.	Boys	44	65	55	16	8/8	15/15	11/11	4/4	38	8.87	100					+.29	+ 6
	Girls	*61*	*55*	*43*	*14*	*11/10*	*11/10*	*9/9*	*3/3*	*34*	*8.58*	*94*						
Defining superior to use.	Boys	44	65	55	16	8/1	15/9	11/4	4/4	38	8.87	47						
	Girls	*61*	*55*	*43*	*14*	*11/2*	*11/2*	*9/5*	*3/0*	*34*	*8.58*	*26*					+.29	+21

The change of the median of the series of differences from —3.5% (non-selected) to 0% (selected) shows that the elimination of over age and special grade pupils has helped the boys more than the girls, and has altered the general relations between the sexes. This fact is also indicated by the average difference in the percentages that all subjects pass each test, the average for non-selected subjects being —1.4% and for selected subjects +1.6%. The non-selected boys from 6 to 12 were given, in all, 2436 tests, these tests being passed 60.8% of the time. The non-selected girls were given 2195 tests, passing 61.6%, the advantage being 0.8% in their favor. The selected boys were given 1125 tests, passing 64.3%, an advantage of 0.1% over the girls who passed 64.2% of 1034 tests. The foregoing changes indicate clearly that the selection of subjects has changed the general relations between the sexes, helping the boys more than the girls.

The relations between the results of selected and non-selected subjects may be studied by a comparison of the differences between the percentages passed by all subjects. If the differences between the scores of the boys and girls are due to but one factor, that of sex differences, then the correlation between the two methods of study should be very nearly absolute. The correlation (Pearson product-moments formula) between the differences in the percentage passed by all boys and girls according to the two methods is 0.726 (pe=0.075). This correlation between the two methods is high, but it would probably be high inasmuch as the 167 selected subjects are included in the 352 non-selected subjects. The results of the two methods show certain large discrepancies. The changes of the greatest magnitude are those shown by the 60 words test (+4% by the first method to + 18% by the second), the tests of defining in terms superior to use (+7% to +21%), of naming the days of the week, (—16% to —2%), giving rhymes, (—10% to +1%), naming colors, (—14% to —4%), copying the diamond, (+1% to —8%), and counting from 20 to 0 (—8% to —16%). The comparison of the median differences shows that the selected method tends to improve the results of the boys more

than those of the girls. All of the changes in the results of the two methods are not in favor of the boys, however, the total scores on the diamond and 20 to 0 tests showing changes in favor of the girls. If the cause of the variations shown by the first method is the presence of a few children of non-English speaking parents, to special class and minus grade children, then the elimination of this source of error should change the results in only one direction.

The analysis of the results of selected subjects, therefore, does not lessen the difficulty of the interpretation of the results in the light of sex differences. The rate of growth of the various abilities with age is irregular. The analysis of the irregularities points to the fact that the boys or girls of any age may be a chance selection of superior or inferior subjects at that age. The method of comparing selected subjects would tend to eliminate the inferior selection of subjects, but would not eliminate the possibility of a superior selection.

The comparison of the results of the sexes shows differences at certain ages and on certain tests that are as high as 20%. The problem involved is that of deciding whether these large differences are due to chance or to differences in the reactions of the sexes. Certain tests show large deviations first in favor of one sex and then in favor of the other. If a difference of a percentage of any magnitude on any test is to be attributed to a sex difference, then the same line of reasoning will show that in certain tests the abilities change from one sex to the other. The analysis of the tests that show this crossing of ability should throw light on the other tests.

Three tests show substantial differences in favor of both sexes according to both methods. In the test of copying the diamond, the non-selected girls lead at the start, age 6, and the boys are ahead at 7, 8 and 9, the same relations being shown by selected subjects of 6 and 8. In the test of copying the designs from memory, the non-selected girls are 24% below the boys at age 9 and 21% above the boys at age 12, the same relations being shown by the selected subjects of 10 and 12. In the test of naming 60 words in three minutes, the non-selected girls are

19% above the boys at 9, and 19% below at 12. The selected boys of 10 and 12 are in advance of the girls in this test.

These three tests are crucial in the consideration of the problem of whether differences shown between the boys and girls are due to actual sex differences or due to accidental causes. Each of these tests may be studied by a method more accurate than that of comparing the percentage passed at each age. The reproductions of the diamond were arbitrarily sorted in six groups according to their merits by a method described in the discussion of the personal equation. The first group contained the best reproductions, the sixth, the poorest. The reproductions of the designs were graded from 0 to 20 by an arbitrary point system described under the discussion of the personal equation. A measure of the ability in the 60 word test is the actual number of words given in three minutes, a measure recorded by the experimenters in each case. Table 16 shows the average score made by the non-selected and selected boys and girls of each age in these three tests.

TABLE 16

Average Score (Mean Variation) of Subjects of Each Age on Three Tests.

		Copying the Diamond Average Group of the Reproductions.		Drawing the Designs Average number of points scored.		Naming 60 words Average number of words given in three minutes.	
		Boys	Girls	Boys	Girls	Boys	Girls
unselected subjects	6	4.27(1.28)	3.57(1.24)				
	7	2.85(1.04)	3.17(1.37)				
	8	2.20(1.15)	3.24(1.57)	8.06(6.19)	9.00(5.25)		
	9	2.33(0.89)	3.00(1.29)	10.29(5.30)	5.32(4.61)	52.93(11.20)	59.91(10.10)
	10			9.17(5.33)	9.18(6.73)	68.12(13.12)	61.76(11.25)
	11			8.64(6.73)	10.94(7.06)	73.65(13.35)	71.28(14.25)
	12			8.64(6.02)	11.08(6.08)	68.75(12.28)	58.14(12.57)
selected subjects	6	4.27(1.20)	3.33(1.26)				
	8	2.32(1.00)	3.00(1.17)				
	10			9.55(5.60)	7.29(6.42)	67.31(12.74)	62.13(11.39)
	12			12.53(5.38)	13.56(5.55)	75.33(10.92)	66.84(13.87)

The relations indicated by the percentage passed are also indicated by the more reliable method of comparing the average scores. In the test of copying the diamond, the 6 year non-selected girls average 0.70 group better than the boys, while the

selected girls are 0.94 ahead. The comparison of the 7, 8 and 9 year subjects shows the boys ahead in all cases, the 8 year non-selected boys averaging over one group higher. The non-selected boys show an improvement of two groups from 6 to 9, while the girls show an improvement of only half a group. One sex shows a decided growth of ability, the other practically none. If the differences indicated are to be taken as real, it will be necessary to assume that the girls pick up the ability to draw a diamond easier than the boys, but that this ability once obtained remains constant—that the effect of maturity operates on one sex but not on another. The number of cases on which this assumption is based (174 subjects from 6 to 9) is so small, and the chances of variation in the selection of subjects of different intellectual status in each age group is so large, that the assumption is not substantiated.

The relations indicated in the test of copying the designs are more variable than those of the diamond test. The 9 year non-selected boys show an improvement over the 8 year boys, but from 9 to 12 there is a gradual decrease in the ability, so that the 11 and 12 year boys are only slightly ahead of the 8 year boys. The relations shown by the non-selected girls are exactly the reverse of those of the boys. The 9 year girls are very much lower than the 8 year girls, and a gradual increase appears from 9 to 12 instead of a decrease. The comparison of these opposite relations gives a maximum difference in favor of the boys at 9 and the girls at 12. If the relations indicated in this test are to be considered definite, the assumption is involved that the influence of increasing age on one sex is exactly opposite to that on the other sex, an assumption that is not substantiated in view of the small number of cases (183 subjects from 8 to 12) and the possibility of selecting subjects of chance superiority in the small groups at each age.

The relations indicated in the test of naming 60 words are more constant than those shown in the diamond or design tests. Both sexes show a growth of ability from 9 to 11 and a decrease from 11 to 12. The growth is irregular, however, the girls showing less growth from 9 to 10, and a greater drop from 11 to

12, so that a comparison of the sexes shows a deviation in favor of the girls at 9 and of the boys at 12. The assumption of any large sex differences in this test involves the assumption that 12 year girls have less ability in this test than 9 year girls, and that the influence of maturity operates differently on the two sexes, an assumption that is not substantiated in view of the many variable factors.

The conclusion that a definite crossing of ability between the sexes occurs in the tests of copying the diamond, copying designs and naming 60 words, is not substantiated. It is not justifiable to attribute a difference of 20% between the sexes to a real sex difference on one test and not on another. If the differences shown between the results of the sexes in the tests of constructing a sentence containing one idea, of naming the months, naming the days of the week, counting stamps and naming colors are to be attributed to sex differences, then the variations in ability shown in the diamond, design and 60 word test must be assumed to be definite. These assumptions were not found to be substantiated, however, so that it is not possible to draw any conclusions concerning sex differences from a study of the percentage that selected or unselected subjects of each age pass each test.

The variable influences due to the selection of subjects of different status at each age are eliminated or counterbalanced to some extent by combining the subjects of all ages. The differences between the percentages that all boys and girls pass each test are to some extent influenced by the ages of the subjects to whom each test was given. The correlation (Pearson product-moments formula) of the differences between the percentages that all non-selected boys and girls passed each test with the difference between the average ages of all the non-selected boys and girls to whom each test was given is 0.394 (pe=0.134). The correlation between the same arrays from selected subjects (i.e. between Columns D and E of Table 15) is 0.388 (pe=0.135). These correlations between the tests and age are high enough to indicate that the factor of age is present to some extent. The close correspondence in the correlations from the two methods

indicates that the age factor is present to the same extent in both methods. The tests vary in the degree with which they correlate with age, so that it is not possible to estimate the amount of the influence of this factor. Furthermore, it has been seen that the results from the two methods are not in strict accordance, that the elimination of inferior subjects caused changes in the results in both directions. For these reasons, it is not possible to draw any conclusions concerning sex differences from a comparison of the percentages passed by all subjects.

Certain negative conclusions are, however, possible. The number of subjects at each age in both methods is comparatively small. The chances of variations due to factors other than sex differences has been shown to be very large. The fact of correspondence between the results of the two sexes is therefore of more importance than the fact of divergence. 75% of the differences between the non-selected boys and girls are 17% or under, while the same proportion of the differences between selected boys and girls falls under 14%. If it is assumed that the subjects of any age should not test lower than those of any preceding age, and allowance is made for differences between the sexes that are exaggerated on account of the chance falling off of ability with older subjects, only 9% of the differences between the non-selected boys and girls are over 20% (derived from Table 13).

The evidence from the foregoing methods of study points to the conclusion that the sex differences, if present, are under 20% or 25% as a maximum, and that deviations of this magnitude are marked exceptions to the general run of differences. The conclusion that the differences that might possibly be attributed to the sex factor are slight, has no meaning unless the word "slight" is defined independently of the writer's personal opinion. The differences shown between the results of the sexes are smaller than those that were attributed to the factor of the personal equation in the study of the results of the four experimenters. It was concluded that certain tests were influenced by grade training. These tests showed from 40% to 60% improvement from one grade to another, so that the greatest influence that may be attributed to the sex factor is only approximately

one half that due to grade training. The following study of the diagnostic value of the tests will show that the deviations that might be attributed to the sex factor are insignificant when compared to the differences between the reactions of normal and retarded children to the individual tests.

Most of the investigators who have studied the factor of sex differences in the Binet tests, have studied them from the standpoint of the "mental ages" or total scores made by the subjects of both sexes. A few investigators have studied sex differences in the light of the individual tests. Descoeudres (20) reports the results of the application of the Binet tests to 24 subjects, one good and one poor pupil of each sex from each of six school grades, drawing conclusions from this investigation concerning the diagnostic value of the individual tests and the sex differences involved. Obviously the number of subjects is too small to allow any conclusions to be drawn. Chotzen (18) compared the percentage that all feeble-minded boys and girls passed each of 15 tests, finding differences varying in magnitude from 1% to 20%. The largest deviations were those of 20% in favor of the boys in the test of copying the diamond, 13% in favor of the girls in the test of executing three commissions, 12% in favor of the boys in naming the pieces of money, 11% in favor of the girls in the test of repeating a sentence of 16 syllables, and 10% in favor of the girls in detecting omissions in pictures. All other differences were less than 10%.

Bloch and Preiss (9) examined 155 normal Volkschule children (79 boys and 76 girls) varying in age from 7 to 13. Bobertag's translation was used. These investigators found very striking differences in the reaction of the sexes to the individual tests, the differences running as high as 52%, most of them in favor of the boys. The differences between the performances of the boys and girls of each age were calculated, without reference to the many sources of variation. The factor of the personal equation is not treated, and this factor alone might cause these variations. If a more careful analysis of the results had been made, it is very probable that the conclusions would have been modified to some extent. The fact that the 11 year

boys are 37% higher than the 11 year girls on the test of criticising absurdities is most certainly modified by the fact that the 11 year subjects are 30% lower than the 10 year subjects in the test of repeating 7 digits. The small number of subjects (in five cases less than 10), would tend to emphasize chance variations. The fact that the number of subjects is too small to warrant definite conclusions is pointed out by the authors. Stern (62) in commenting upon these results, points out the significance of the fact that the inferiority of the girls extends to so many different kinds of tests. The results of Bloch and Preiss are in almost complete contradiction to the results of the present investigation. They find large differences, and find practically all of these differences in favor of the boys. This investigation shows a general run of differences very much smaller, and a slight general superiority of the non-selected girls. The mere fact of contradiction in the results of the two investigations would indicate that the differences were not produced by the common factor of sex. Rogers and McIntyre (54) give no figures, but report that they have studied their results in the light of sex differences, and have found no correlation between their results and those of Bloch and Preiss.

The results of the investigators who have compared the "mental ages" or total scores of children of different sexes are somewhat at variance. Goddard (30) reports that there are more backward boys than girls. Stern notes that Goddard's results do not bear out his statement, for the percentage of boys and girls testing two or more years retarded is the same (18.5%). The accuracy of Goddard's statement depends on the criterion[5] used for measuring backwardness. Although Goddard's state-

[5] If the criterion is four or more years retarded, there are more backward boys than girls (boys = 3.7%, girls = 3.1%). If the criterion is three or more years backward, there are more girls than boys (boys = 8%, girls = 9.1%). If the criterion is two or more years backward, the proportions are the same, as Stern notes. If the criterion is one year or more retarded, there are more backward boys than girls (boys = 41.4%, girls = 35.6%). There are more girls than boys testing at and above age according to Goddard's results. 34.7% of the boys and 36.6% of the girls test at age, while 23.8% of the boys and 27.7% of the girls test one year or more above age.

ment concerning the backwardness of the boys may be interpreted differently, his figures leave no doubt concerning the fact that there are more girls than boys at and above age, and therefore indicate a general superiority of the girls.

Bobertag (10) computed the average "mental age" of 90 boys and 90 girls regularly distributed from 7 to 12. The subjects were selected according to school grades, so that the average grade of each group differed by exactly one grade. His results show the boys ahead 0.06 yr. at 7, 0.14 yr. at 8 and 9, 0.20 yr. at 10, 0.19 yr. at 11 and 0.14 yr. at 12. These findings cannot be considered entirely out of harmony with those of Goddard, for, as this investigation shows, there may be a change in the relation of non-selected boys and girls and selected boys and girls.

Yerkes and his co-workers (82), scoring some of the Binet tests according to the point system, show that the girls of English speaking parents are superior to the boys of the same parentage between 5 and 7, that they fall below with minor variations till 11, where they again surpass the boys at 12 and 13, falling below at 14 and 15. The differences between the sexes are smaller and of less practical importance than the differences due to the language factor, but the authors suspect "that at certain ages serious injustice will be done to individuals by evaluating their scores in the light of norms which do not take account of sex differences." (page 73). In contradiction to these results are those of Terman and his co-workers (67), who, scoring the Stanford revision of the Binet scale according to "intelligence quotients," find differences of but 2% to 4% in these quotients in favor of the girls, and who conclude from the basis of their studies of sex differences that the conclusions of Yerkes are unjustified. These two investigations used tests different in character and differently weighted, so that the results would not necessarily have to correspond.

The one common feature of most of the researches on sex differences in the Binet-Simon tests is that the differences are small. Burt and Moore (17) summarize the work of various investigators in the general field of sex differences, and report an investigation of their own on 67 boys and 63 girls, 12½ to 13½

years of age. They discuss their results and those of the other authors in the order of the complexity of the mental processes involved. They find a high correlation between the size of the sex difference and the simplicity of the capacities compared—the higher the process, and the more complex the capacity, the smaller the sex difference.

The general trend of the investigations on sex differences indicates that no very large differences are to be expected in the application of intelligence tests, and that the differences to be expected will vary according to the nature of the tests. The results of this investigation are in agreement with the general trend of the investigations in showing only slight differences that might be attributed to the sex factor. The results do not show on what tests, if any, these differences occur. Conclusions concerning the amount of influence of this factor must be drawn from more exhaustive investigations on the individual tests. The research of Bateman (3), for instance, is conclusive in the test of naming colors. Bateman shows that there is a difference of 14% in favor of the girls in this test, showing furthermore that the factor of school training causes an improvement of but 18%. The results would indicate that the test should be placed in the fifth or sixth year, but the sex difference of 14% would probably not warrant the placing of the test in a different age group for boys and girls.

The investigations of Bolton (11) and Wooley (79) would show that small differences in favor of the girls are to be expected in the tests of repeating digits, and possibly in all memory tests. The investigations of Gilbert (27), Thompson (68), Burt and Moore, and Peterson and Doll (51) would indicate that a slight difference in favor of the boys should appear in the test of arranging five weights. Ruger's (55) finding of striking differences in favor of men in a series of puzzle tests, and Wooley and Fisher's finding of large differences in favor of the boys in the Healy puzzle-box test would show that rather large differences might appear in the general class of "puzzle" tests.

Even though the sex differences in intelligence tests may be shown to be small, scientific procedure should demand that the

investigator who standardizes any test or system of tests should treat his results in such a way as to demonstrate that the factor is present or not present. The burden of proof should still be on the person who maintains that sex differences are not involved. The knowledge of sex differences is especially important in diagnosing border-line cases of mental defect, where the diagnosis must often be made on the qualitatively different character of the responses to individual tests.

VI. SUMMARY.

One of the fundamental assumptions in the construction of the Binet-Simon scale is the correlation of the individual tests with age. The correlation of the tests with age is affected by the error due to incomplete data, by the influence of the personal equation of the experimenter, and by the training the subject has received in school.

The influence of the personal equation of the experimenter was found to be more marked in some tests than in others, the influence being most marked in the tests of copying the diamond, indicating omissions in pictures, defining in terms superior to use, drawing designs from memory, detecting absurdities in statements and reconstructing dissected sentences.

The variations between the experimenters could be traced to three sources,—

1) to the use of apparatus, variations in which were due to,
   a) the construction of the test material, and
   b) the use of alternative questions;

2) to the technique of the experimenters in giving the tests; and

3) to observation errors made by the experimenters in marking a response passed or failed.

It is possible to eliminate all three sources of error.

The effect of school training was more marked on some tests than on others, the effect being most marked in the tests of counting stamps, counting backward from 20 to 0, enumerating the days of the week and the months, giving the day and the date, naming the pieces of money, making change, and reconstructing dissected sentences. Tests that involve school training should be standardized on a different basis than those relatively independent of this factor.

Although the comparison of "mental ages" and pedagogical ages gives no information concerning the general correlation be-

tween the Binet tests and the school grades, the study of the individual tests establishes the fact of a general correlation.

The correlation of the individual tests with grade is higher than the correlation of the tests with age, this fact being indirect evidence of the value of the tests as measures of intelligence.

Sex differences were found to be slight as compared with the influence due to the personal equation or grade training.

Since variations occur in the results due to the influence of the personal equation and grade training, certain allowances must be made for these factors in making diagnoses on the basis of the tests. The scale is therefore a qualitative rather than a quantitative instrument.

The investigator who wishes to use his results for standardizing age norms should use only those data based on the complete method of experimenting, and should treat his results in such a way as to demonstrate the presence or absence of the variable factors of the personal equation, grade training and sex differences.

BIBLIOGRAPHY

1. Abelson, A. R. The Measurement of Mental Ability of "Backward" Children. Brit. J. of Psychol., 1911, 4, 268-314.
2. Ayres, L. P. The Binet-Simon Measuring Scale of Intelligence: Some Criticisms and Suggestions. Psychol. Clinic, 1911, 5, 187-196.
3. Bateman, W. G. The Naming of Colors by Children. Ped. Sem., 1915, 22, 469-486.
4. Binet, A. Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'ecole. Année psychol., 1911, 17, 145-201.
5. Binet, A. and Simon T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. Année psychol., 1905, 11, 191-244.
6. Binet, A. and Simon T. Application des méthodes nouvelles au diagnostic du niveau intellectual chez des enfants normaux et anormaux d'hospice et d'ecole primaire. Année psychol., 1905, 11, 245-336.
7. Binet, A. and Simon T. Le développement de l'intelligence chez les enfants. Année psychol., 1908, 14, 1-94.
8. Binet, A. and Simon T. La mesure du développement de l'intelligence chez les jeunes enfants. Bull. de la soc. libre pour l'etude psychol. de l'enfant. 1911, 11, 187-256.
9. Bloch, E. and Preiss, A. Ueber intelligenzprüfungen an normalen Volksschulkindern nach Bobertag. (Methode von Binet und Simon) Zsch. f. agnew. Psychol., 1912, 6, 539-547.
10. Bobertag, O. Ueber Intelligenzprüfungen (nach der Methode von Binet und Simon). I. Methodik und Ergebnisse der einzelnen Tests. Zsch, f. angew. Psychol., 1911, 5, 105-203. II. Gesamtergebnisse der Methode. Zsch. f. angew. Psychol., 1912, 6, 495-537.
11. Bolton, T. L. The Growth of Memory in School Children. Amer. J. of Psychol., 1892, 4, 362-380.
12. Bonser, F. G. The Reasoning Ability of Children of the Fourth, Fifth and Sixth School Grades. New York: Columbia Univ., 1910, pp. 133.

13. BRIDGMAN, O. Mental Deficiency and Delinquency. J. of Amer. Med. Assoc., 1913, 61, 471-472.
14. BRIGHAM, C. C. An Experimental Critique of the Binet-Simon Scale. J. of Educ. Psychol., 1914, 5, 439-448.
15. Buffalo conference. J. C. Bell, C. S. Berry, W. S. Cornell, E. A. Doll, J. E. W. Wallin, G. M. Whipple, Informal Conference on the Binet-Simon Scale: Some Suggestions and Recommendations. J. of Educ. Psychol., 1914, 5, 95-100.
16. BURT, C. Experimental Tests of General Intelligence. Brit. J. of Psychol., 1910, 3, 94-177.
17. BURT, C. AND MOORE, R. G. The Mental Differences between the Sexes. J. of Exp. Ped., 1912, 1, 273-284, 355-388.
18. CHOTZEN, F. Die Intelligenzprüfungsmethode von Binet-Simon bei schwachsinnigen Kindern. Zsch. f. angew. Psychol., 1912, 6, 411-494.
19. DECROLY, O. AND DEGAND J. La mesure de l'intelligence chez des enfants normaux d'après les tests de M. Binet et Simon: nouvelle contribution critique. Arch. de psychol., 1910, 9, 81-108.
20. DESCOEUDRES, A. Les tests de Binet et Simon et leur valeur scolaire. Arch. de psychol., 1911, 11, 331-350.
21. DESCOEUDRES, A. Exploration de quelques tests d'intelligence chez des enfants anormaux et arriérés. Année psychol., 1911, 11, 351-375.
22. DOLL, E. A. Inexpert Binet Examiners and their Limitations. J. of Educ. Psychol., 1913, 4, 607-609.
23. DOUGHERTY, M. L. Report on the Binet-Simon Tests given to 483 Children in the Public Schools of Kansas City, Kansas. J. of Educ. Psychol., 1913, 4, 338-352.
24. DRESSLAR, F. B. Studies in .the Psychology of Touch. Amer. J. of Psychol., 1894, 6, 313-368.
25. EBBINGHAUS, H. Ueber eine neue Methode zur Prüfung geistigen Fahigkeiten und ihre Andwendung bei Schulkindern. Zsch. f. Psychol. 1897, 13, 401-459.
26. FERNALD, W. E. The Diagnosis of the Higher Grades of Mental Defect. Amer. J. of Insan., 1914, 70, 741-752.
27. GILBERT, J. A. Researches on the Mental and Physical Development of School Children. Stud. fr. Yale Psychol. Lab., 1894, 2, 40-100.
28. GODDARD, H. H. The Binet-Simon Measuring Scale for Intelligence. (Revised edition) Vineland, N. J. The Training School, 1911, pp. 16.

29. GODDARD, H. H. Standard Method of giving the Binet Test. Training School, 1913, 10, 23-32.
30. GODDARD, H. H. Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence. Ped. Sem., 1911, 18, 232-259.
31. GODDARD, H. H. Three Annual Testings of 400 Feeble-Minded Children and 500 Normal Children. Psychol. Bull. 1913, 10, 75-77.
32. HAINES, T. H. Diagnostic Value of some Performance Tests. Psychol. Rev., 1915, 22, 299-305.
33. HEALY, W. The Individual Delinquent. Boston: Little Brown & Co., pp. 830.
34. HEALY, W. AND FERNALD, G. M. Tests for Practical Mental Classification. Psychol. Monog. 1911, 13 (No. 54) pp. 53.
35. HUEY, E. B. The Binet Scale for Measuring Intelligence and Retardation. J. of Educ. Psychol., 1910, 1, 435-444.
36. HUEY, E. B. A Point Scale of Tests for Intelligence. Baltimore: Warwick & York (folder) 4 pp.
37. KATZENELLENBOGEN, E. W. A Critical Essay on Mental Tests in their Relation to Epilepsy. Epilepsia, 1913, 4, 130-173.
38. KITE, E. S. The Binet-Simon Measuring Scale of Intelligence. Philadelphia: Committee on Provision for the Feeble-Minded, Bull. no. 1, pp. 29.
39. KITE, E. S. The Development of Intelligence in Children. (Contains translations of nos. 5, 6, and 7). Vineland, N. J.: The Training School (Publications of the Department of Research, No. 11), 1916, pp. 328
40. KITE, E. S. The Intelligence of the Feeble-Minded. (Translation of three articles by Binet and Simon on Feeble-mindedness) Vineland, N. J.: The Training School, (Publications of the Department of Research, No. 12), 1916, pp. 328.
41. KOHS, S. C. The Binet-Simon Measuring Scale of Intelligence: an Annotated Bibliography. J. of Educ. Psychol., 1914, 5, 215-224, 279-290. 335-346.
42. KOHS, S. C. The Practicability of the Binet Scale and the Question of the Borderline Case. Training School, 1916, 12, 211-224.
43. KUHLMAN, F. Some Results of Examining a Thousand Public School Children with a Revision of the Binet-Simon Tests of Intelligence by Untrained Examiners. J. of Psycho-Asthenics, 1914, 18, 233-269.

44. MARTIN, A. L. A Contribution to the Standardization of the De Sanctis Tests. Training School, 1916, 13, 93-110.
45. MEUMANN, E. Vorlesungen zur Einführung in die experimentelle Pädagogik und ihre psychologischen Grundlagen. Leipzig: W. Englemann 1913, Vol. II, pp. 800.
46. MEUMANN, E. Ueber eine neue Methode der Intelligenzprüfung und über den Wert der Kombinationsmethoden. Zsch. f. päd. Psychol. und exp. Päd., 1912, 13, 145-163.
47. MORROW, L. AND BRIDGMAN, O. Delinquent Girls Tested by the Binet Scale. Training School, 1912, 9, 33-36.
48. NORSWORTHY, N. The Psychology of Mentally Deficient Children. New York: (Columbia Univ. thesis) 1906, pp. 111.
49. OTIS, A. S. Some Logical Aspects of the Binet Scale. Psychol. Rev. 1916, 23, 129-152, 165-179.
50. OTIS, M. The Binet Tests Applied to Delinquent Girls. Psychol. Clinic, 1913, 7, 127-134.
51. PETERSON, A. M. AND DOLL, E. A. Sensory Discrimination in Normal and Feeble-Minded Children. Training School, 1914, 11, 110-118, 135-144.
52. PILLSBURY, W. B. The Psychology of Reasoning. New York: D. Appleton & Co., 1910, pp. 304.
53. PYLE, W. H. A Psychological Study of Bright and Dull Pupils. J. of Educ. Psychol., 1915, 6, 151-156.
54. ROGERS, A. L. AND MCINTYRE, J. L. The Measurement of Intelligence in Children by the Binet-Simon Scale. Brit. J. of Psychol., 1915, 7, 265-299.
55. RUGER, H. A. Sex Differences in the Solution of Mechanical Puzzles. (In report of New York branch of American Psychological Assoc.) J. of Phil., Psychol., etc., 1914, 11, 412-413.
56. SCHMITT, C. The Binet-Simon Tests of Mental Ability. Ped. Sem. 1912, 19, 186-200.
57. SCHMITT, C. Standardization of Tests for Defective Children. Psychol. Monog., 1915, 19 (No. 83) pp. 181.
58. SIMPSON, B. R. Correlations of Mental Ability. New York: Columbia Univ., 1912, pp. 122.
59. SMITH, F. O. The Effect of Training in Pitch Discrimination. Univ. Iowa Stud. in Psychol., Vol. VI. Psychol., Monog., 1914, 16 (No. 69) 67-103.
60. STENQUIST, J. L., THORNDIKE, E. L. AND TRABUE, M. R. The Intellectual Status of Children who are Public Charges. Arch. of Psychol. 1915. 33, pp. 52.

61. STERN, W. Die differentielle Psychologie in ihren methodischen Grundlagen. Leipzig: Barth, 1911, pp. 503.
62. STERN, W. The Psychological Methods of Testing Intelligence. (Whipple, G. M., trans. fr. German) Educ. Psychol. Monog., No. 13, Baltimore: Warwick & York, 1914, pp. 160.
63. Symposium on Mental Tests. (Conducted by C. E. Seashore under "Communications and Discussions") J. of Educ. Psychol., 1916, 7. (R. M. Yerkes, 163-164).
64. TERMAN, L. M. Genius and Stupidity. Ped. Sem., 1906, 13, 307-373.
65. TERMAN, L. M. The Measurement of Intelligence. Boston: Houghton Mifflin Co., 1916, pp. 362.
66. TERMAN, L. M. AND CHILDS, H. G. A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence. J. of Educ. Psychol., 1912, 3, 61-74, 133- 143, 198-208, 277-289.
67. TERMAN, L. M., LYMAN, G., ORDAHL, G., ORDAHL, L., GALBREATH, N. AND TALBOT, W. The Stanford Revision of the Binet-Simon Scale, and some Results from its Application to One Thousand Non-Selected Children. J. of Educ. Psychol., 1915, 6, 551-562.
68. THOMPSON, H. B. Psychological Norms in Men and Women. Chicago: Univ. of Chicago Press, 1903, pp. 188.
69. THORNDIKE, E. L. The Significance of the Binet Mental Ages. Psychol. Clinic, 1914, 8, 185-189.
70. THORNDIKE, E. L. An Introduction to the Theory of Mental and Social Measurements. New York: Teachers' College, 1913, pp. 277.
71. THORNDIKE, E. L., LAY W. AND DEAN, P. R. The Relation of Accuracy in Sensory Discrimination to General Intelligence. Amer. J. of Psychol., 1909, 20, 364-369.
72. TOWN, C. H. A Method of Measuring the Development of The Intelligence of Young Children. (Authorized translation of no. 8) Lincoln, Ill.; Courier-Herald Co. 1913, pp. 82.
73. WALLIN, J. E. W. Experimental Studies of Mental Defectives. Educ. Psychol. Monog. No. 7. Baltimore, Warwick & York, 1912, pp. 155.
74. WITMER, L. On the Relation of Intelligence to Efficiency. Psychol. Clinic, 1915, 9, 61-86.
75. WHIPPLE, G. M. Manual of Mental and Physical Tests. Baltimore: Warwick & York, 1910, pp. 534.

76. WHIPPLE, G. M. Manual of Mental and Physical Tests. Baltimore: Warwick & York, 1914, pp. 690, 2 vol.
77. W[HIPPLE], G. M. The Amateur and the Binet-Simon Tests. J. of Educ. Psychol., 1912, 3, 118-119.
78. W[HIPPLE], G. M. Amateruism in Binet Testing once more. J. of Educ. Psychol., 1913, 4, 301-302.
79. WOOLEY, H. T. A New Scale of Mental and Physical Measurements for Adolescents and some of its Uses. J. of Educ. Psychol. 1915, 6, 521-550.
80. WOOLEY, H. T. AND FISHER, C. R. Mental and Physical Measurements of Working Children. Psychol. Monog. 1914, 18 (No. 77) pp. 247.
81. WYATT, S. The Quantitative Investigation of Higher Mental Processes. Brit J. of Psychol., 1914, 6, 109-133.
82. YERKES, R. M., BRIDGES, J. W. AND HARDWICK, R. S. A Point Scale of Measuring Mental Ability. Baltimore: Warwick & York, 1915, pp. 213.

THIS BOOK IS DUE ON THE LAST DATE
STAMPED BELOW

AN INITIAL FINE OF 25 CENTS
WILL BE ASSESSED FOR FAILURE TO RETURN THIS BOOK ON THE DATE DUE. THE PENALTY WILL INCREASE TO 50 CENTS ON THE FOURTH DAY AND TO $1.00 ON THE SEVENTH DAY OVERDUE.

JUL 29 1943

REC'D LD

MAR - 1 1953

LD 21-100m-7,'39 (402s)

YD 22588

UNIVERSITY OF CALIFORNIA LIBRARY

studied. The use of the normal grade age as a measure of scholastic ability is false inasmuch as it rests on the assumption that all children enter school at a certain age, which is not the case. The measure of scholastic ability is the measure of the child's reaction to the subject matter of the grades, and that measure may be expressed only in the fact of promotion, non-promotion or (very rarely) double promotion, in other words, it may be expressed only in the relation of grade to the length of time in school. Furthermore, the two measures of scholastic ability, the age in grade method, and the grade progress method, are measures of an historically past performance not of present possibilities, and the true measure of an ability must indicate potential ability.

As measures of scholastic ability in terms of actual reaction, these measures present a distribution of general ability that is skewed toward the lower end, or in the direction of no ability. If a child enters school late, he presents a picture of retardation according to the age and grade method, while through any number of causes independent of intellectual ability, a child may present a retardation of at least a year according to either method. The possibilities for advancement are not as great, however, for advancement means forcing a child through a mass of subject matter, a process which the school is generally unwilling to undertake and the parent is generally unwilling to sanction. The school therefore presents a picture of ability in which promotion is normal, and non-promotion far more frequent than advance. If general ability is to be considered as distributed over any sort of a frequency surface, that surface will not take the form presented by the school measure in which the modal ability is almost completely the upper limit.

The measure of "mental age" has been shown to be one which varies from one chronological age to another in the form of its distribution. Normal children of 6 or 7 test over age, while those of 11 and 12 test under age. This abnormal distribution is due to two facts. In the first place, the tests in the younger years are too easy and those in the higher years are too difficult.

In the second place, the younger children have a wider range of tests beyond their average ability, so that exceptional subjects may display exceptional ability in a manner that is impossible if ability is measured by school progress, while older children have only a few tests within their range, the picture of advancement being excluded as in the measure of school ability. If the mental ages of a run of subjects of different chronological ages are combined, the frequency surface is normal, the error of the extremities balancing.

The investigators who have compared "mental age" with grade age, have compared two distributions, one of which is markedly skewed, the other normal, but false. The resulting finding of mental advance in excess of pedagogical advance has significance only insofar as it shows that a measure of general ability that will admit of exceptionally high performance is a better measure than one that precludes the possibility of such performance. The only significant finding is that pupils who show marked retardation in school rarely if ever show mental advance.

Applying the foregoing discussion to Schmitt's results in particular, all that has been said concerning the inadequacy of the age in grade method applies to her results. The age for entering school being 5, none of the subjects in the Kindergarten could be advanced, while those who entered late would be retarded. It is difficult to see how these young children would be able to make up their work in such a way as to show advance during the first two or three school years. The normal age for the sixth grade is from 11.5 to 12.5 years. Inasmuch as no grades were tested above VI, none of the 37 subjects from 11.5 to 14.5 could show an advance, and all of the 19 subjects from 12.5 to 14.5 would necessarily show retardation. Schmitt's results differ from those of other investigators in finding more subjects advanced according to Binet age in relation to chronological age. This deviation is probably due to the fact that she examined a superior selection of subjects, and to the fact that the XV year and "Adult" tests were used, so that the older subjects, who in general fall below their chronological age, had an opportunity to

better their scores. The discrepancy shown by Schmitt between school standing and the Binet tests does not demonstrate the inadequacy of the tests.

The final demonstration of a correlation between the Binet scale and school grade, rests not in comparing the total score or "mental age" with school grade, for that is susceptible to the errors of over-estimation and under-estimation according to varying chronological age, but in comparing the results of subjects in each grade on the individual tests. The tests may vary in their correlation with grade. Inasmuch as there is a general growth in age with grade, and a corresponding growth of intelligence with age, a test, in order to be an adequate test of intelligence, must show a correlation with grade. If the correlation is too high, however, the value of the individual test is in question for it would then be testing, not intelligence, but grade training. This criterion was actually used, though not stated, by Binet in his discussion of the results of Decroly and Degand (19), and in his revision of the 1908 scale, in which many of the tests that he considered to relate to school training were eliminated.

Studies of the individual tests in the light of school grade are not available. Decroly and Degand published in 1910 the results of an investigation on 45 children in a Brussels school, similar in character to that studied by Schmitt in Chicago. Binet discussed these results and those of other minor investigations in the Paris schools in considering the effect of environment on the results of the tests. Although he referred to school training as a factor, and classified the tests in which Decroly and Degand's subjects were superior, he gave no quantitative demonstration of the effect of this factor. The results of Decroly and Degand are based on too few subjects to admit of quantitative treatment. Chotzen (18) studied the tests by comparing the performance of feeble-minded individuals of the same mental age but of different chronological age. Although this method shows the effect of environment and maturity on feeble-minded individuals, it does not bear directly on the factor of school training. The foregoing

investigations will be discussed in this chapter only in their relation to the results of the particular tests. Schmitt, in her monograph, published tables showing the reaction of each subject in each grade to each test, the tables being discussed in the text. Although it was not Schmitt's purpose to determine the correlations between the various tests and grade, her data are available for a study of this sort, and the writer has taken the liberty of figuring them in this light, indicating at the same time Schmitt's interpretation of the grade factor, contained in the accompanying text. These data will be compared with the results of the Princeton investigation.

422 subjects of this investigation were distributed in the kindergarten, first six regular grades, minus grades and the special class of the Princeton Model School. 301 of the subjects (161 boys and 140 girls) were in the kindergarten and first six regular grades. The data obtained from the examination of these 301 subjects were classified according to the grade in which the subjects were found, and the percentage that the subjects of each grade passed each test was calculated.

Only those tests were studied which showed themselves to be free from the influence of the personal equation of the four experimenters. The elimination of the unrecorded results of the definitions test left a number of cases too small to be studied. To avoid the influence of the error due to incomplete data, the writer has calculated the percentage from only those tests that were given from 75% to 100% of the possible number of times. The data from the tests of repeating 5, 6 and 7 digits have been combined into one weighted measure. The procedure of the experimenters in giving these tests was to start within the subject's range and continue till he failed. If 5 digits were successfully repeated, 6 were given, and if these were passed, 7 were given. The results have been combined into one measure for the sake of simplicity, 1 point being allowed for the successful repetition of 5 digits, 2 points for 6 digits and 3 points for 7 digits, the weighting being roughly in accordance with the weighting in Goddard's scale, the tests being in the age groups

VIII, X and XII respectively. The measure of the ability of a group to repeat digits is the per cent. that the number of points scored is of the number of points possible (i.e. 6 times the number of subjects in the group).

The number of subjects in each grade (boys and girls shown separately) the average age of the subjects in each grade, together with the mean variation from the average are shown in Table 3.

TABLE 3

Number of Boys and Girls in Each Grade, and the Average Age of All Subjects in Each Grade.

Grade	Number of Boys	Number of Girls	Total No. of Subjects	Average Age	Mean Variation
Kindergarten	20	12	32	5.64 years	0.46 years
Grade I	27	24	51	7.05 "	0.50 "
Grade II	16	24	40	8.16 "	0.65 "
Grade III	21	24	45	9.31 "	0.75 "
Grade IV	20	15	35	10.46 "	0.91 "
Grade V	24	25	49	11.71 "	0.99 "
Grade VI	33	16	49	12.81 "	1.06 "

The above table shows an increase of a year or more (actually from 1.10 years to 1.41 years) in the average age of the subjects in each grade. From this it is reasonable to expect that there is a general growth in intelligence correlating with this increase in age, or, in other words, to expect a correlation between the results of the individual tests and the grade in which the performance occurred. If the correlation is too high, it will indicate a dependence of that particular test on the subject matter of the grade. In Table 4 are shown the percentages that the subjects in each grade passed each test. The notes referred to in the margin contain the proportions passed for all other subjects for whom the percentages are not given, the percentages being given only for those groups to whom the tests were given from 75% to 100% of the possible number of times.

A study of Table 4 shows that the tests in general correlate with grade. The combined score of the test of repeating digits, for example, shows a growth from 6% to 78%, more rapid in the first three grades than in the last four. The tests vary in

TABLE 4

Percentage that Subjects in Each Grade Passed Each Test. 301 Subjects.

Test	Grades K	I	II	III	IV	V	VI	
VII-1, 13 pennies	72	96	100					Note 1
VII-2, Pictures	69	96	94					Note 2
VII-4, Diamond	46	75	88					Note 3
VII-5, Colors	72	90	97					Note 1
VIII-2, 20 to 0		9	53	80				Note 4
VIII-4, Stamps		13	50	78				Note 5
All digits, (combined)....	6	21	42	51	55	78	75	
VIII-3, Days of week....	16	45	90	100				Note 6
IX-3, Date		5	35	96	100			Note 7
IX-4, Months			28	84	90			Note 8
X-I, Money			20	36	57	82		Note 9
X-2, Designs				21	37	42	66	Note 10
X-5, Sentence (2 ideas)..				67	89	88	98	Note 11
XI-2, Sentence (1 idea)....				22	46	51	74	Note 12
XI-3, 60 words					63	63	87	Note 13
XI-4, Rhymes					67	63	76	Note 14

Note 1. Counting 13 pennies and naming colors given 20 times above II. Not failed.
Note 2. Describing pictures given 21 times above II. Not failed.
Note 3. Copying diamond given 25 times above II. Not failed.
Note 4. Counting from 20 to 0 given 18 times in K. Not passed. Given 31 times above III. Failed once.
Note 5. Counting stamps given 15 times in K. Not passed. Given 35 times above III. Failed 3 times.
Note 6. Naming days of week. Given 32 times above III. Not failed.
Note 7. Giving day and date given 5 times in K. Not passed. Given 56 times above IV. Not failed.
Note 8. Naming months. Given 26 times below II. Passed twice. Given 44 times above IV. Failed twice.
Note 9. Naming money. Given 26 times below II. Passed 3 times. Given 28 times in VI. Failed twice.
Note 10. Copying designs given 33 times below III. Passed 5 times.
Note 11. Sentence (2 ideas) given 32 times below III. Passed 12 times.
Note 12. Sentence (1 idea) given 32 times below III. Passed 4 times.
Note 13. Giving 60 words given 53 times below IV. Passed 19 times.
Note 14. Giving rhymes given 42 times below IV. Passed 26 times.

the number of grades taken to reach their maximum. The test of naming the day and date, for example, is failed by all subjects in the kindergarten, 95% of Grade I and 65% of Grade II, while only 4% of the subjects in Grade III and none of those in the

higher grades fail it. A sudden increase occurs between Grades II and III showing possibly the influence of grade training. The tests vary considerably in the degree of their correlation. An easily obtained measure of the degree of correlation is that of comparing the magnitude of the increases from grade to grade. For example, there is an increase of 61% (96%—35%) from Grade II to Grade III in the ability to pass the test of giving the day and date, and an increase of 16% (36—20%) between the same grades in the test of naming the pieces of money. The former test correlates higher with the influence of grade in this particular case than the latter.

In this manner the percentage difference between the performance of the subjects in each grade and that of the subjects in the preceding grade was obtained. All the increases or decreases in ability from one grade to another were thus obtained, these values serving as measures of the amount of correlation between the tests and the grades. 42 differences between the performance of the subjects in any grade and those of the next succeeding grade were thus obtained. In 4 cases there were actual decreases of 1, 2, 3 and 4% which were not significant. The difference ranged from —4% to +61%, the median being +19.5% (Q=16.25%). Some of the differences between the grades might be due to the chance superiority of a particular grade. To overcome this chance variation, and to furnish another index of the growth of the various abilities, the differences were calculated by steps of two grades, i.e., subtracting the performance of the kindergarten from the second grade, the first from the third, etc. In this way, 26 differences were obtained varying from +9% to +91%, the median being +29% (Q=18%).

Some of the differences noted are undoubtedly high enough to warrant the assumption of the effect of grade training on the tests. Just what tests show this effect is probably a matter of opinion. Allowance must be made for the growth of an ability independent of training. 25% of the highest increases from one grade to another were selected as being worthy of special consideration at least. A larger increase must be allowed be-

tween two grades. Those differences were considered worthy of special consideration that exceeded twice the value of the median of the one-grade differences or 39%. This manner of selecting the largest differences is quite arbitrary, but is justified by the outcome, for the tests that show the most significant increases according to this method show those increases in more than one step, so that the evidence is concentrated against a very few tests. In this way the significant values outweigh the less significant values and fair allowance is made for growth from one grade to another.

The following list includes the tests showing the greatest increases by one-grade and two-grade steps, together with the magnitude of the increases and the grades between which they occur.

One-grade steps. 25% of largest increases.	Two-grade steps. Increases greater than 39%.
+61% Date, II to III	+91% Date, I to III
+56% Months, II to III	+74% Days, K to II
	+71% 20 to 0, I to III
+45% Days, I to II	
+44% 20 to 0, I to II	+65% Stamps, I to III
	+65% Date, II to IV
+37% Stamps, I to II	+62% Months, II to IV
+30% Date, I to II	
	+55% Days, I to III
+29% Diamond, K to I	
+29% Days, K to I	+46% Money, III to IV
+28% Stamps, II to III	+42% Diamond, K to II
+27% 20 to 0, II to III	
+27% Pictures, K to I	

The above lists of increases are confined to but 8 tests. In all, there were 16 tests studied. According to the method of selecting the significant increases, 20 such values actually appeared. In this manner the evidence combines against a very few tests. Some tests appear in both lists and more than once in the same list. The most striking growth with grade is shown in the tests of giving the day and date, naming the months, nam-

ing the days of the week, counting from 20 to 0 and counting stamps. The tests of copying the diamond, describing pictures and naming money may or may not show this influence. The evidence is strongest in the case of the diamond test since that appears in both lists.

The foregoing method of selecting those tests which correlate with grade to such an extent as to indicate the influence of grade training is not conclusive, owing to the fact that there is also an increase in age from grade to grade. If a test showed a very rapid growth with age, and those ages fell for the most part in certain grades, then those grades would show an increase which might be wrongly assumed to be due to training. The tests of counting from 20 to 0 is a case in point. Yerkes (82) in Table 32, page 125, gives the percentage values for each test in the Point Scale, for English speaking boys and girls of each age. The test, of the twenty one tests included, that shows the most marked increase with age is that of counting backward, the values being as follows,— age 4=0%; age 5=3.5%; age 6=23.7%; age 7=45.7%; age 8=72.2%; age 9=96%; the values for ages above 9 being 97% or higher.

The age in grade distribution of the 301 subjects in this investigation is given in Table 5.

TABLE 5

Distribution of Subjects in Each Grade according to Chronological Age.

	Grades							
Age	K	I	II	III	IV	V	VI	Total
4	4							4
5	17							17
6	11	28	2					41
7		18	17	2	1			38
8		4	15	18	1			38
9			5	13	11			29
10		1		10	14	18	1	44
11			1	2	3	16	16	38
12					5	8	11	24
13						4	12	16
14						2	5	7
15						1	3	4
16							1	1
Total	32	51	40	45	35	49	49	301

The rapid growth of the ability in counting from 20 to 0, according to the method of comparing the subjects in each grade, was from 9% in Grade I to 80% in Grade III. From Table 5 it may be seen that practically all, (89%), of the chronological ages in Grades I, II and III were distributed in the ages 6, 7, 8 and 9, a chronological range coinciding with that in which Yerkes' results show the ability to develop. The growth of this ability might be due then either to age or to grade. For this reason, to arrive at any final conclusion, it is necessary to compare the subjects of the same age but in different grades. The treatment of the Princeton results according to this method follows, but the analysis of the data in this manner can have no great reliability owing to the small number of subjects in each group. The number of subjects in each group, (boys and girls shown separately), the average age and mean variation from this average are shown in Table 6.

TABLE 6

Number of Boys and Girls of Similar Ages in Different Grades, and the Average Age of the Subjects of Similar Ages in Each Grade.

Grade	Age	Number of Boys	Number of Girls	Total no. of Subjects	Average Age	Mean Variation
Kindergarten.	5	11	6	17	5.48	0.20
Kindergarten.	6	8	3	11	6.26	0.21
Grade I	6	14	14	28	6.59	0.17
Grade I	7	9	9	18	7.36	0.22
Grade II	7	7	10	17	7.56	0.24
Grade II	8	6	9	15	8.39	0.24
Grade III ...	8	8	10	18	8.60	0.22
Grade III ...	9	5	8	13	9.43	0.16
Grade IV	9	5	6	11	9.65	0.13
Grade IV	10	10	4	14	10.39	0.30
Grade V	10	7	11	18	10.54	0.25
Grade V	11	10	6	16	11.54	0.22
Grade VI ...	11	10	6	16	11.53	0.26
Grade VI	12	6	5	11	12.52	0.14

All chronological ages were computed in tenths of a year, so that a variation in age from 0.1 yr. to 0.9 yr. is possible within

TABLE 7

Actual percentage that each test was passed by the subjects of each age in each grade. 223 subjects.

AGE	Age 5	Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12
GRADE	K	K	I	I	II	II	III	III	IV	IV	V	V	VI	VI
NO. OF SUBJECTS	17	11	28	18	17	15	18	13	11	14	18	16	16	11
Counting 13 pennies	71	91	96	94	100	100	Note 1							
Describing pictures	65	82	96	100	100	86	Note 1							
Copying diamond	50*	45	64	89	93	92	Note 2							
Naming colors	71	82	89	100	100	100	Note 1							
Counting from 20 to 0			16	0	63	43	83*	Note 3						
Counting stamps			13	12	29	60	85*	Note 4						
Repeating digits, (all)	6	6	21	19	42	46	47	51	65	48	79	83	76	86
Naming days of week	0	40	50	50	94	80	100*	Note 5						
Giving the day and date					29	20	88	100	100	100	Note 6			
Naming months					25	27	88	85	100	Note 7				
Naming pieces of money						31	35	31	55	64	83	83	Note 8	
Copying designs							31	31	27	43	41	57	Note 9	
3 words in sentence. 2 ideas					Note 10		89	62	91	93	94	75	93	100
3 words in sentence. 1 idea					Note 11		33	8	45	50	61	50	80	80
60 words in 3 minutes						Note 12			44	77	50	88	93	80
Giving rhymes						Note 13				92	56	69	87	100

*The test of copying the diamond was given 59% of the possible number of times in K-5, counting from 20 to 0 and giving days of week, 66% in III-8, and counting stamps 72% in III-8. All other percentages are based on tests given from 75 to 100% of the possible number of times.

Note 1. Tests of counting 13 pennies, describing pictures and naming colors each given 12 times above II-8. No failures.
Note 2. Copying diamond given 15 times above II-8. No failures.
Note 3. Counting from 20 to 0 given 16 times below I-6. Not passed. Given 31 times above III-8. Failed 4 times.
Note 4. Counting stamps given 14 times below I-6. Not passed. Given 32 times above III-8. Failed 4 times.
Note 5. Giving days of week given 32 times above III-8. No failures.
Note 6. Giving date given 39 times below II-7. Passed twice. Given 36 times above IV-10. No failures.
Note 7. Naming months given 24 times below II-7. Passed twice. Given 37 times above IV-9. Failed 4 times.
Note 8. Naming pieces of money given 35 times below II-8. Passed 4 times. Given 14 times above V-11. Failed twice.
Note 9. Copying designs given 26 times below III-8. Passed 5 times. Given 15 times above V-11. Failed 6 times.
Note 10. Three words in sentence, 2 ideas, given 24 times below III-8. Passed 9 times.
Note 11. Sentence, 1 idea, given same as 2. Passed 3 times.
Note 12. 60 words in 3 minutes given 41 times below IV-9. Passed 10 times.
Note 13. Giving rhymes given 37 times below IV-10. Passed 25 times.

each age group. That the subjects of the "same" age but in different grades are not exactly the same is shown in Table 6. The subjects of each age in the higher grades average from 0.01 yr. to 0.33 yr. different, with an average superiority of 0.19 yr. This difference, however, is about one fourth that between the subjects or different ages in the same grades, and may be called the same for practical purposes. For convenience, the groups will be referred to as K-5, II-7 etc., the first member referring to the grade, the second to the age. K-5 would mean the group of 5 year children in the kindergarten, II-7, the 7 year subjects in Grade II, etc. The actual per cent, that the subjects in each group passed each test was calculated and is shown in Table 7. Unless otherwise noted, the percentages are based on tests given 75% to 100% of the possible number of times.

Some of the groups from which results were obtained are too small to have great reliability, but the method is at least suggestive. The results of 14 groups are given. It is possible then to compare the results of subjects of 6 ages, (6, 7, 8, 9, 10 and 11), that are in different grades, and also to compare sub-

jects in all seven grades that are of different ages, and in this way to determine whether the dominating factor in the growth of any ability is that of grade or age. The reliability of the method rests only on its connection with that of the first method employed.

In answer to the question of whether the growth of ability in the test of counting from 20 to 0 is due to age or grade, a question which was unanswered by the first method, we may turn to the results shown in Table 7 in which the subjects of each age in each grade are shown. The test of counting from 20 to 0 was not passed by any of the 5 and 6 year subjects in the kindergarten. Comparing first the subjects of different ages in the same grade, the 7 year subjects in Grade 1 are 16% lower than the 6 year subjects in that grade, and the 8 year subjects in Grade II are 20% lower than the 7 year subjects in the same grade, the older subjects making a lower record in each case. Comparing the performance of the subjects of the same age but in different grades, the 7 year subjects in Grade II are 63% ahead of the subjects of the same age in Grade I, while the 8 year subjects are 40% [1] ahead of the subjects of the same age in Grade II. Allowing for the retrogression of the older subjects in each group, i.e. assuming that they should have done equally as well as the younger subjects in the same grade, the groups in Grades II and III are still 47% and 20% ahead of the subjects in the grades lower. The growth of ability in this test would therefore appear to be due to grade training.

A rapid growth of ability in the test of counting stamps occurred between Grades I and III (37% I-II+28% II-III=65% I-III), according to the first method, so that the same question arises as in the test of counting from 20 to 0. The test was not passed below group I-6. No growth with age is shown between

[1] This test was given to but 66% of the subjects in III-8, the experimenters assuming that the other 34% would pass. The score given, 85%, therefore represents the ability of the lowest selection of III-8 subjects, or the most conservative estimate of the ability of the whole group. The same applies to the other tests in III-8 given 66% and 72% of the time. In this way the hypothesis that the tests are not influenced by grade training is given the benefit of the doubt.

I-6 and I-7, but a growth of 31% appears between II-7 and II-8. A growth with grade of 17% is shown from I-7 to II-7 and of 25% from II-8 to III-8. This test shows therefore the operation of the two factors of age and grade training.

The improvement in ability in the tests of counting 13 pennies, describing pictures and naming colors, that was indicated between the kindergarten and Grade I by the first method, would refer to age rather than grade, for a greater increase in each test is indicated between K-5 and K-6 than between K-6 and I-6. Above I-6 these abilities are completely developed. It could be maintained that these tests are so completely within the ability of the groups that the effect of training would not be indicated. The test that is best adapted to show the influence of any factor on a group is one that is well within the ability of the group—the influence of the factor will be obscured if the measure is either too easy or too difficult. The test of copying the diamond is a case in point and one well worth study, for it has been attributed to the effect of training by various authors. All the reproductions of the diamond had been scored according to the arbitrary system outlined in the previous discussion of the personal equation. A control on the factor of difficulty was obtained by raising or lowering the passing mark in this test. The percentage passed was calculated for each group for each of the 5 possible passing marks. The relations indicated in Table 7, where the passing mark is Group IV, were not changed by this process of raising or lowering the passing mark. In all cases the influence of age was shown between groups I-6 and I-7, and the influence of grade shown between groups K-6 and I-6. The test was given to but 59% of the K-5 group, the experimenter assuming that the other 41% would fail, so that the percentages calculated represent the performance of the best selection of K-5 subjects, or, in other words, the benefit of the doubt is given to the hypothesis that the test is influenced by grade training. If the other members of K-5 had failed according to the experimenter's assumption, (and this assumption was quite justified for some had failed to draw the square), 29% of the group would have passed instead of 50%.

The influence of age indicated in this test is as great if not greater than that due to training.

The test of repeating digits, scored by the weighting system previously described, exhibits a slow but uniform progress throughout, the older subjects in each group making records that are about the same or slightly lower than those of the younger subjects in the same grade, an increase showing fairly regularly from grade to grade. The most marked increase in this ability appears between K-6 and I-6, and between I-7 and II-7, possibly indicating that the lack of familiarity with the use of digits in the lowest grades interferes with this test as a measure of auditory memory.

The test of naming the days of the week shows the most marked improvement with age (40%) from K-5 to K-6, practically no improvement (10%), from K-6 to I-6, no improvement from I-6 to I-7, a very marked increase with grade from I-7 to II-7, a drop from II-7 to II-8, group III-8 marking the complete development of the ability. The test would appear to be due to the combined effect of age and grade. The tests of giving the day and date and naming the months are passed only twice in the kindergarten and first grade, by about a quarter of the subjects in II-7 and II-8 without age increase, while the subjects in III-8 shows a most marked increase due to grade. Above III-8 these tests are seldom failed. The test of naming the pieces of money shows a slow growth from 8 to 11, the largest increases appearing from III-9 to IV-9 and from IV-10 to V-10, improvement with grade in each case. Copying the designs from memory shows a growth of 26% from 8 to 11, the development occurring in two age steps, from IV-9 to IV-10 and from V-10 to V-11.

The growth with age cannot be determined in the tests of constructing sentences from three given words, because they were given to too few cases below the third grade. The results do not show whether III-8 is exceptionally high or III-9 exceptionally low. Both tests show decreases in ability from III-8 to III-9 and from V-10 to V-11. The ability in the easier test is well within the range of the third and higher grades, showing, therefore, no

improvement. The improvement in the second test develops from 33% to 80% in three steps, correlating with Grades IV, V and VI in each case. The most vital question, that of determining whether or not the language training in the third grade helps to make the construction of a sentence possible, cannot be determined owing to the lack of material in the second grade. The experimenters' assumptions in not trying the test would indicate this fact, but this is not experiment. The same lack of material makes conclusions in regard to the rhyming test impossible. The performance of IV-10 is exceeded only by VI-11. The test of naming 60 words in three minutes shows two decided increases with age and one decided drop with grade.

The foregoing analysis is based on a number of subjects in each group too small to have any great significance. The general fact of the correlation of the tests with grade remains, and conclusions concerning what tests correlate too highly with training can be answered only by considering both methods of study, and by considering only the largest deviations. The two most striking instances are found in the tests of naming the months and giving the date. These tests undoubtedly relate almost entirely to training. Less striking but equally definite is the relation of the test of counting from 20 to 0 to training. The tests of naming the days of the week and counting stamps show the influence of age to an extent almost as marked as that of grade, so that while the development in these tests is rapid, the grade factor probably exerts only part of the influence. Conclusions concerning the other tests are largely a matter of opinion, and the opinion of the writer has been indicated in the detailed discussion.

A study of the tests in relation to grade by the first method employed may be made from Schmitt's results. The author gives, in Table I, II, III, IV, V, VI and VII on pages 70, 71, 73, 74, 75, 76 and 77 of her monograph, the results of each subject in each grade on each test. From these tables the present writer has calculated the percentage passed in each test. A study of this sort rests for its reliability on the accuracy of the published tables, and the facts indicated by the tables do not always coincide

with Schmitt's discussion.[2] The writer has followed the tables rather than the discussion in calculating the results. In the VIII-2 test where an alternative rank is given for counting from 10 to 0 instead of 20 to 0, the writer has considered success in counting from 10 to 0 as a failure in counting from 20 to 0. In the line suggestion test Schmitt recognizes two types of failure, the typical failure according to Binet of accepting the suggestion of the first three lines, and the failure due to the fact that the subject actually judges the lines unequal after studying them. The second type of response Schmitt marks as passed, using a special symbol. The writer has calculated these percentages separately, entering the first or Binet type of response under "Line suggestion A" in the table, and the second type under "B." The V year and Adult tests were omitted. All of the other tests were included that had been given over 70% of the possible number of times. Unless otherwise noted, each test was given 100% of the possible number of times. Table 8 shows the per cent. that Schmitt's subjects in each grade passed each test in Binet's 1911 scale (Town's translation with modifications). The table is given with the reservation that the tables from which the percentages were calculated might contain misprints, and that the writer's interpretation of the tables might be at fault.

Inasmuch as there are many differences in procedure in giving the tests, and in the character of the schools tested, the results of the two investigations are not comparable in respect to the percentage passed in one grade in one study with those in the same grade in the other study. The method used in determining the

[2] In the discussion (page 69) Schmitt gives 15 subjects in the kindergarten failing test VII-4. Table I shows 13. On the same page she gives 24 subjects failing VIII-4. Table I shows 22 failing. In discussing the results of Grade I (page 72) Schmitt states that there is "more than 50% of failure with the discrimination of weight", while Table II shows 35% failure. Again, the tests referred to specific school instruction by Schmitt are VII-4, VIII-4, and IX 1, 2, 3 and 4. On page 72, in discussing the results of Grade I, she says "the tests below ten years which depend upon specific instruction are usually not passed except the VII-4 test. The percentages passed are as follows: VII-4 = 85%; VIII-4 = 45%; IX-1 = 35%; IX-2 = 75%; IX-3=90%; IX-4=30%. "Usually not passed" includes, therefore, tests passed 75% and 90% of the time.

TABLE 8

Per cent. that Schmitt's Subjects of Each Grade Passed Each Test. 150 Subjects.

	Grades						
	K	I	II	III	IV	V	VI
Number of subjects	25	20	17	21	22	22	23
VI-1, Distinguishing morning, afternoon	96	100*					
2, Defining in terms of use	92	94*					
3, Copying diamond	76	94*					
4, Counting 13 pennies	92	100*					
5, Choosing prettier of faces	92	100*					
VII-1, Showing right hand	92	80	100				
2, Describing pictures	72	65	81				
3, Executing 3 commissions	92	95	100				
4, Counting stamps	48	85	100				
5, Naming colors	96	100	100				
VIII-1, Comparing remembered objects	92	100	100	100	100		
2, Counting backwards from 20 to 0	40	85	94	95	100		
3, Indicating omissions in pictures	100	95	94	100	100		
4, Giving day and date	12	45	94	100	100		
5, Repeating 5 digits	64	85	94	100	100		
IX-1, Making change	6*	35	71	95	86	100	
2, Defining in terms superior to use	39*	75	65	100	95	100	
3, Naming pieces of money	28*	90	94	100	100	100	
4, Naming the months	6*	30	71	95	95	95	
5, Comprehending easy questions	61*	100	100	95	100	100	
X-1, Arranging 5 weights		65	41	57	50	64	
2, Copying designs		10	35	57	45	32	
3, Detecting absurdities		60	88	100	100	100	
4, Comprehending difficult questions		85	100	100	100	100	
5, Constructing sentence. Two ideas		65	76	100	100	100	
XII-1, Resisting suggestion, A. (Binet scoring)		64*	76	52	41	14*	100
B. Judgment error counted plus				100	86	100*	
2, Constructing sentence. One idea		57*	71	95	95	100*	100
3, Giving 60 words in three minutes		43*	82	62	100	95*	96
4, Defining abstract terms		7*	29	52	73	95*	100
5, Reconstructing dissected sentences		0*	6	10	23	81*	78
XV-1, Repeating 7 digits						62*	78
2, Rhyming words with "obey"						86*	70
3, Repeating a sentence of 26 syllables						10*	17
4, Interpreting pictures						14*	70
5, Solving problems from various facts						62*	70

Note.—All tests except those marked (*) were given all the possible number of times. The VI year tests were given 90% of the time in Grade I, the IX year tests 72% of the time in the kindergarten, the XII year tests 70% of the time in Grade I, and the XII and XV year tests 95% of the time in Grade V.

correlation of the tests with grade is the same as that used in the first method of treating the Princeton data, that of comparing the differences between grades by one-grade and two-grade steps, of selecting an arbitrary standard for detecting exceptional growth, and of comparing the resulting lists. The differences between the performance of each grade and the next succeeding grade were calculated. These differences, 100 in number, ranged from —24% to +62%, the median being +5% (Q=10.75%). The run of differences differs from that found in the Princeton study in two respects, in having a lower median and variability, and in containing more minus deviations. The lower median and variability is due to the fact that the tests were given over a wider range, the Princeton tests being given only on the "up slope" of the growth curve, or not being given when the tests were any distance above or below the probable range of ability of the group. The Princeton results showed only 4 minus deviations of 4, 3, 2, and 1% respectively, while Schmitt's results show 15 such deviations, 6 of them being 10% or over. These deviations are probably due to the smaller number of subjects, and if due to chance, should be counteracted by the precautionary measure of combining the indices of correlation into two-grade steps. 71 two-grade differences were obtained ranging from — 25% to + 82%, the median being + 10% (Q=16.5%). 4 measures were still in the minus direction, one of these, —25% (Design III to V) is probably significant, the other values of —6%, —5% and —4% having no significance. Inasmuch as the variability of the series is lower, those differences were considered to be worthy of special study that had the value of 2Q+M, or were in excess of the interquartile range plus the median. The lists of tests that appear as showing marked growth with grade according to the two methods are as follows:

One grade differences higher than 2Q+M	Two grade differences higher than 2Q+M
+62%, IX-3, Money, K to I	+82%, VIII-4, Date, K to II
+58%, XII-5, Dissected, IV to V	+71%, XII-5, Dissected, III to V
+49%, VIII-4, Date, I to II	+66%, IX-3, Money, K to II
+45%, VIII-2, 20 to 0, K to I	+65%, IX-4, Months, K to II
+41%, IX-4, Months, I to II	+65%, IX-4, Months, I to III
+39%, IX-5, Comprehension, K to I	+65%, IX-1, Change, K to II
+39%, XII-3, 60 words, I to II	+60%, IX-1, Change, I to III
+38%, XII-3, 60 words, III to IV	+55%, XII-5, Dissected, IV to VI
+37%, VII-4, Stamps, K to I	+55%, VIII-4, Date, I to III
+36%, IX-2, Definitions, K to I	+54%, VIII-2, 20 to 0, K to II
+36%, IX-1, Change, I to II	+52%, VII-4, Stamps, K to II
+35%, IX-2, Definitions, II to III	+47%, X-2, Design, I to III
+33%, VIII-4, Date, K to I	+45%, XII-4, Abstract Def., I to III
+29%, IX-1, Change, K-I	+44%, XII-4, Abstract Def., II to IV
+28%, X-3, Absurdities, I to II	+43%, XII-4, Abstract Def., III to V

A study of the above lists shows, as in the similar study of the Princeton data, that although the method of selecting the exceptional tests is an arbitrary one, the method is justified in practice, for only a few tests (13) appear in the lists as significant. In all, there were 34 tests[3] studied, and 30 differences were considered large enough to be significant. These 30 differences were confined to 13 tests. The tests of naming 60 words and defining in terms of use drop out of the first list owing to the elimination of the errors of negative correlation. The design test is both positive and negative, the ability increasing from Grades I to III and decreasing after III. The test of defining abstract terms appears according to the second method because the ability increases with grade from 7% in I to 95% in V by

[3] No differences were calculated from the line suggestion test owing to the possibility of misinterpreting the symbols. Schmitt notes the difference in the character of the responses from the suggestion error to the judgment error in passing from Grade II to III. The scoring of the suggestion error in the tables shows an inverse correlation with Grades II, III, IV and V, and a sudden change again from 14% in Grade V to 100% in Grade VI, so that there is probably a mistake. The scoring of the responses to this test according to the strict Binet ruling would make the "mental ages" lower, for many cases would then have basal X.

increases of approximately 25% in each grade. No conclusions may be drawn concerning the easy comprehension test and the absurdities test. The 20 remaining differences are confined to 7 tests, those of naming the day and date, naming the months, counting from 20 to 0, counting stamps, naming money, reconstructing dissected sentences, and making change. The first four were included in the five found to show the most marked influence of grade in the Princeton study. The test of naming the pieces of money did not show a marked relation to grade in the latter study, but this difference might be one of school curriculum. The test of naming the days of the week is not included in Binet's 1911 scale.

In the Princeton study alternatives were used in the making change question so that no data from this test were included in the quantitative study. These data show the ability in this test developing in the second and third grades, the test being passed only twice in the kindergarten and first grades, and generally passed above the third. The data in the test of reconstructing dissected sentences show very few passing the test below grade V with approximately three fourths passing in V and VI. In so far as the Trenton experimenting was applied to a few subjects in the regular grades below the seventh, this test was rarely passed in the third and fourth grade, passed about 5% in V, and almost universally passed in VI, VII and VIII. The number of subjects in each grade is small in the Trenton experiment, but each test was separately scored, i.e. each part of the dissected sentence test, each part of the absurdity test etc. Each of the three parts of the dissected sentence test showed the same growth between the same grades, and this growth was more marked than that in any other test. The evidence concerning these two tests, therefore, supports the evidence from Schmitt's results.

The quantitative analysis of the Princeton data and Schmitt's data would indicate that the tests of counting stamps, counting from 20 to 0, naming the days of the week, giving the day and date, naming the months, naming the pieces of money, making change and reconstructing dissected sentences were influenced to a considerable extent by grade training. The performance in

certain of these tests (days, date and months) may be the result of specific school training in the tests themselves, while others (perhaps the tests of counting stamps, counting from 20 to 0, and reconstructing dissected sentences) may involve a transfer effect in the application of the content of the grade in a new way. The fact that the tests correlate very highly with grade training does not show that the tests are worthless, but it does show that they should, perhaps, be placed in another scale, or should at least be placed on a different footing than those that test capacity irrespective of attainments.

One of the best tests[4] of intelligence is the determination of what an individual can do with the training he has received, but tests of this sort rest on the assumption that the individual's opportunities have been determined. The importance of tests of information in cases of alienation presenting a picture of deterioration is recognized. The important change to be made is not the elimination of such tests from intelligence scales, but their standardization on a different basis. The diagnostic value of such tests rests not in the mechanical memorizing of a time series such as that of the months, but in the ability to apply such a series. In pointing out this fact Katzenellenbogen (37) suggests that the months test be given in some such manner as "If somebody asks you in November to return three months later, what month would it be?" Decroly and Degand also suggest that the mechanical tests of counting and naming the days of the week and months be modified in some such manner.

[4] The writer recalls two cases in which the failure in tests which involved the application of training was very significant. The first was that of a woman of about 30, a parole patient in a hospital for the insane, who had never shown any marked symptoms other than a history of intellectual inferiority. This patient passed practically all of the Binet tests in the IX, X and XII year groups, but failed completely in the test of making change. This observation was later checked up. Another case of a woman of 22, in the same hospital, presented a border-line psychoneurotic picture perhaps, but no marked symptoms other than a history of intellectual inferiority. She passed in a great many of the difficult tests in the upper years but had great difficulty in telling time. Both cases had lived under very good home conditions and had mingled with people of ability. A great many tests of capacity were given, but the most illuminating evidence of their mental status came from the two tests mentioned.

Comparing the conclusions of this study with other investigations, the agreement is fairly close. Schmitt's results do not support her suggestion that the definitions test relates to specific school instruction. The other tests which she refers to this factor (stamps, date, 20 to 0, change, months and money) show the influence to a marked extent. Binet in classifying some of the tests referred the tests of copying a sentence, reading for memories, writing from dictation, copying a diamond, counting backwards and making change to scholastic training. The first three tests were not included in this investigation. The diamond test showed the influence of age to be as great if not greater than that of school training. The last two tests showed a marked influence of training. Binet referred the tests of counting 13 pennies, naming four colors, naming the days of the week and enumerating the months to home training. The last two showed a marked influence of school training. The results of the present investigation agree with those of Chotzen in finding no effect or very little effect of training in the tests of copying the diamond. repeating digits, describing pictures, counting 13 pennies, naming colors, comparing remembered objects, defining in terms of use and superior to use, and in finding marked influence of this factor in the test of naming the days of the week.

The methods used in analysing the results, especially the second method, reveal several suggestive relations between the tests and the school grades. There is a general correlation between the tests and the grades, a correlation that is very necessary to establish, for there is also a general correlation between intelligence and grade. In analysing the results of the individual tests by comparing the results of subjects of the same age in different grades, and of subjects of different ages in the same grade (Table 7), it was seen that, as a general rule, the growth in any particular ability occurred in passing from grade to grade, not in passing from age to age within one grade. In fact in only half of the cases in which the subjects of two ages in one grade may be compared do the older subjects make records that are higher than those of the younger ones, and only 10% of these gains are over 20%. If the groups were considered to be equal in all

cases in which their records were within 10% of each other. equality occurs in exactly 50% of the cases. Of the remainder, 20% of the groups were lower, while in only 30% of the cases are the older subjects actually higher than the younger subjects of the same grade. Some of the cases of retrogression could well be accidental, but they occur too frequently to be due entirely to chance.

Applying the same general method to the cases in which groups of the same age but in different grades were compared, 5% of the groups in a higher grade showed lower scores, the results correspond in 43% of the cases, while 52% showed definite improvement. This might indicate that there is a higher correlation between the tests and grade than between the tests and age. The fact that the comparison of children of different ages in the same grade showed the older children making lower records in 20% of the cases, equal records in 50% of the cases and higher records in only 30%, would confirm the general diagnostic value of the tests if Bonser's interpretation of this phenomena is correct. Bonser (12) applied various sorts of reasoning tests to children in the fourth, fifth and sixth school grades. In summarizing the results of the tests in the different grades, he says, "In the contrast with grade progress and progress with age, in the generally superior showing made by the younger groups of children of any grade when contrasted with the older pupils of the grade, and in the fairly substantial percentage of pupils from lower grades found in the highest quartile of ability for all, it is shown that native capacity is measured to a high degree by the tests."

In conclusion, the results shown in this chapter would indicate a correlation between the individual tests studied and the school grades, this correlation being high enough in some cases to show the actual effect of training. In answer to the general objection that since one demonstration of the accuracy of the tests rests on their correlation with school grades, the school grades are the real measure of intelligence and the mental tests superfluous, it is only necessary to point out that intelligence tests, besides affording the opportunity for accurate standardization,

also detect the subject's potential abilities independent of his past performance. The school measure indicates mental defect in cases of gross retardation, but it does not indicate exceptional ability.

Schmitt's contention that the school represents a standard environmental situation, and a measure of a subject's ability should include a measure of the adequacy of his reaction to this situation, is well founded. It is not, however, a criticism of the Binet scale, for the scale aims to test native capacity. At the Buffalo conference (15) on the Binet scale, the following question was raised,—"What is it, after all, that the scale aims to test?" The question was answered by "We believe that current misconceptions as to the aim of the scale should be removed. It is not intended to test the emotional or volitional nature, but primarily intelligence (judgment)." To this list might be added the assertion that the scale was not intended to test a child's reaction to the school situation, or to furnish an outline for taking a record of his life history.

Rogers and McIntyre (54) would also have mental tests include tests dependent on both school and home training. This general trend of present day discussion is a reversion to Binet's 1908 type of scale, a tendency to which Binet was in opposition. The probable solution rests in eliminating from the scale the tests involving training, and in constructing a standardized scale of another sort for the estimation of the individual's reaction to the school situation in terms of the length of time that he has met that situation. That such a scale is not a matter of speculation is shown by the number of scales now on the market for measuring handwriting, spelling, composition, arithmetical ability, etc. Tests of native capacity and tests dependent on school and environmental training cannot be standardized on the same basis, for they are essentially different measures. Measures of the first sort may perhaps be correlated with age, while measures of the other sort can be correlated only with opportunity.

V. SEX DIFFERENCES

The investigators who have studied the influence of sex differences on the Binet-Simon tests have used two methods, that of comparing the "mental ages" or total scores of subjects of each sex, and that of comparing the per cent. that the subjects of each sex pass each test. The first method throws no light on the individual tests, inasmuch as one sex may be superior in one test and inferior in another so that the total score will balance the influence of this factor. Inasmuch as the scale is founded on the principle that sex differences do not exist, it is important to study the individual tests, and to determine the accuracy of this assumption.

The Princeton data are available for a study of this sort. 352 subjects (187 boys and 165 girls) between the ages 6 and 12 were examined. The method of study adopted was that of comparing the results of non-selected boys and girls of each age, and, as a check on this method, of comparing the results of selected boys and girls of four ages.

Inasmuch as the subjects of each chronological age are distributed over a range of one year (the 6 year subjects for example being distributed from 6.0 to 6.9), the actual average age of the subjects of each age was computed to make sure that no differences might appear due to the chance selection of subjects at either extreme. These averages are shown in Table 9.

TABLE 9

Actual Average Chronological Age of Boys and Girls in Each Age Group.

	BOYS		GIRLS	
	Number of Subjects	Average Age (M. V.)	Number of Subjects	Average Age (M. V.)
Age 6	37	6.58 (0.20)	23	6.51 (0.20)
Age 7	29	7.50 (0.29)	31	7.39 (0.26)
Age 8	24	8.48 (0.29)	28	8.48 (0.22)
Age 9	20	9.46 (0.27)	22	9.54 (0.26)
Age 10	31	10.46 (0.25)	23	10.37 (0.30)
Age 11	28	11.59 (0.22)	20	11.52 (0.27)
Age 12	18	12.43 (0.30)	18	12.57 (0.24)

for girls 0.83 yr. (from 9 to 10), while the maximum increase for boys is 1.13 yr. (from 10 to 11), and for girls 1.15 yr. (from 10 to 11). A more marked lack of regularity in the growth of scholastic ability from year to year as measured by the average grade is shown in Table 11, no increase being shown by the 12 year boys over the 11 year boys, while the 10 year boys show an increase of 1.44 to 1.01 grades over the 9 year boys. In the same way the 10 year girls show an increase over the 9 year girls that is nearly three times that of the 7 year girls over the 6 year girls, while the increase of the 7 year girls over the 6 year girls is twice that of the 12 year girls over the 11 year girls. These relations indicate that the selection of subjects is not uniform at each age. The subjects of any one age may be either a superior or inferior selection of all children of that age, and there is no reason for supposing that this random sample of superior or inferior subjects of any age will correspond to a similar sampling of the subjects of the opposite sex of the same age.

The process of calculating the percentage that the boys and girls of each age pass each test is extremely simple, but the conclusion, that the differences found between the percentage passed by the sexes at each age may be attributed to sex differences, is not justified unless all the variable factors are known.

A previous chapter showed variations in the tests due to the influence of the personal equation of the experimenters. To avoid this variable influence, only those tests were studied that showed that they were free from the influence of this factor. Inasmuch as each experimenter examined approximately the same number of boys and girls of each age, any influence of this factor would be equalized, provided, of course, that there were no differences in the reaction of the experimenters to the two sexes. In the detailed study of the design test, it was found that experimenter C was more lenient in marking girls than boys. The possibility of a similar interpretation in a few other tests was suggested, but not demonstrated. In analysing the results for sex differences, however, the possibility of such an interpretation must be kept in mind.

Another possible source of error is that due to incomplete data.

The experimenters, in giving the tests, would give only those within the approximate range of the subject, so that each test would be given to a superior selection of children below the normal range of the test, and to an inferior selection of subjects above this range, a process tending to make the apparent growth of an ability less than the probable real growth. In comparing the results of the sexes, however, it is not necessary to have accurate results on the growth of an ability, but results which have the same determining factors. If the experimenters gave the test to approximately the same proportions of boys and girls at each age, a comparison of the percentage passed is legitimate, even if a small proportion of the whole group were actually tested, for the proportion would include the same selection of subjects. The number of boys and girls at each age, and the percentage that each test was given to these subjects are shown in Table 12. The test of counting 13 pennies, for example, was given 37 times to 6 year boys, or 100% of the possible number of times, while the test of counting from 20 to 0 was given 27 times to the same group, or 73% of the possible number of times. Column A shows the total number of times each test was given to all of the boys and girls. Column B gives the average age of all the boys and girls to whom each test was given. The average given in this case is not the actual average derived from the actual chronological age of each subject figured in tenths, but the weighted[1] average, the whole numbers 6, 7, 8, 9, 10, 11, and 12 being used.

Table 12 shows a very close correspondence between the percentage that each test was given to boys and girls of each age, so that the error due to incomplete data, though present, is present to the same extent in the results of both sexes, and may be disregarded. A fairly close correspondence in the average age of all the boys and girls to whom each test was given is also indicated in Table 12. In the test of counting stamps there is an

[1] For example, in the test of counting 13 pennies, the average age of the boys to whom the test was given is,—

$$\frac{(37\times6)+(28\times7)+(16\times8)+(8\times9)+(7\times10)+(3\times11)+(1\times12)}{100} = 7.33 \text{ years}$$

Neither method, then, is entirely satisfactory, the first because it would tend to exaggerate chance differences, the second because it would tend to obscure real differences. The method used in this study is that of comparing the results of non-selected and selected subjects of each age and sex, studying first the general growth of each ability from age to age within each sex, and using the per cent. that all subjects pass each test to determine the correlation between the results of non-selected and selected subjects.

Table 13 shows the percentage of proportion[3] that the boys and girls of each age pass each test, the percentage that all boys and girls pass each test, the actual percentage that the boys are superior to (+) or inferior to (—) the girls of each age, the difference between the average age of all boys and girls to whom each test was given, and the difference between the percentage that all boys and girls pass each test.

The differences between the performance of the boys and girls at each age have no meaning unless the general growth of the abilities in each sex is first understood. Studying first the results of the 187 non-selected boys shown in the first seven columns of Table 13, it may be seen that the growth of ability in each test is rather irregular. The test of naming the months, for example, shows a slight decrease from 9 to 12. The differences between the percentage performances of the subjects of each age and those of the preceding age were calculated. The 12 year group, compared to the 11 year group, is +11% on the test of giving the date, —9% on the test of naming the months etc. 61 differences were thus obtained, varying in magnitude from —15% to +36%, the median being +8% (Q=9.75%). 13 of the deviations (21%) were minus values. The largest negative deviations occurred in the tests of naming colors (—15%, 7 to 8), naming money (—15%, 11 to 12), and constructing a sentence containing two ideas (—13%, 8 to 9). The remaining 10 minus deviations were less than 10%.

[3] The proportion given is the number of times a test was given over the number of times a test was passed. No percentages were calculated for tests given less than 12 times, and no percentages are given for the definitions tests on account of the small number of times they are given to all subjects.

TABLE 13

Percentage or Proportion that the Boys and Girls of Each Age Passed Each Test, and the Differences between the Percentages Passed.

		Chronological Ages							% passed all subjects	Differences in percentage passed at each age							Difference in average ages	Difference in % passed by all
		6	7	8	9	10	11	12		6	7	8	9	10	11	12		
Counting 13 pennies	Boys	84	93	94	8/7	7/7	3/3	1/1	90	−16	− 4	− 6					−.23	− 9
	Girls	100	97	100	9/9	6/6	2/2	2/2	99									
Describing pictures	Boys	86	100	94	9/9	8/8	3/2	1/1	93	− 1	+17	+10					−.28	+ 5
	Girls	87	83	84	9/9	7/7	4/4	2/2	88									
Copying a diamond	Boys	57	93	93	92	10/9	4/2	3/3	79	−21	+17	+22	+13				−.44	+ 1
	Girls	78	76	71	79	8/7	4/4	2/2	78									
Naming four colors	Boys	70	96	81	9/9	7/7	3/1	1/1	83	−30	+ 3	−19					−.27	−14
	Girls	100	93	100	9/8	7/7	3/3	2/2	97									
Counting from 20 to 0	Boys	4	25	55	50	79	6/3	8/7	42	−25	− 7	+14	− 9	+ 4			−.07	− 8
	Girls	29	32	41	59	75	7/6	5/5	50									
Counting stamps	Boys	25	29	62	75	79	6/3	8/7	52	+14	+ 7	+ 7	− 3	+21			0	+ 6
	Girls	11	22	55	78	58	7/5	7/6	46									
Repeating all digits	Boys	12	28	35	56	59	63	73	43	− 9	− 7	− 9	+ 9	− 6	− 7	+ 1	−.03	− 6
	Girls	21	35	44	47	65	70	72	49									
Naming days of week	Boys	35	59	75	88	95	6/5	8/8	67	−24	−16	− 8	−12	− 5			−.06	−16
	Girls	59	75	83	100	100	7/7	5/5	83									
Giving day and date	Boys	13	18	48	84	92	89	100	64	− 4	− 7	− 6	− 7	− 8	−11	0	−.60	− 1
	Girls	17	25	54	91	100	100	100	65									
Naming the months	Boys	7	17	42	74	76	73	64	51		+ 2	−23	+ 6	−18	−19	−36	+.06	−13
	Girls	9/2	15	65	68	94	92	100	64									
Naming pieces of money	Boys	10/2	11	38	33	66	72	57	44		+11	+17	0	− 1	+ 3	−18	+.10	+ 4
	Girls	10/1	0	21	33	61	69	75	40									
Copying designs from memory	Boys	6/0	9/3	25	35	24	36	29	28			0	+24	−12	− 8	−21	+.10	0
	Girls	6/0	6/0	25	11	36	44	50	28									
3 words in sent. Two ideas	Boys	3/1	9/3	80	67	74	73	78	70			+17	+ 4	+ 9	−11	−10	+.43	+ 3
	Girls	6/0	8/4	63	63	65	84	88	67									
3 words in sent. One idea	Boys	3/1	9/3	27	22	45	58	67	44			+ 6	+ 1	+15	+21	+26	+.43	+18
	Girls	6/0	8/0	21	21	30	37	41	26									
Naming 60 words in 3 minutes	Boys	4/0	6/3	9/4	36	60	77	69	58				−19	+ 7	− 6	+19	+.17	+ 4
	Girls	7/0	3/0	9/6	55	53	83	50	54									
Giving rhymes with 3 words	Boys	3/0	6/4	6/5	10/4	57	60	65	58					− 2	− 1	−15	+.16	−10
	Girls	3/1	4/3	10/8	10/7	59	61	80	68									
Defining by use	Boys	14/13	18/16	12/12	11/10	15/15	7/6	3/3	94								+.26	+ 4
	Girls	14/13	20/15	17/16	15/14	9/9	4/4	2/2	90									
Defining superior to use	Boys	14/1	18/6	12/7	11/6	15/3	7/2	3/1	33								+.26	+ 7
	Girls	14/3	20/4	17/5	15/5	9/3	4/1	2/0	26									

An index of the growth from year to year was obtained by calculating the average percentage increase from one age group to another. For example, the 7 year boys were 26% higher than the 6 year boys in the test of naming colors, 5% higher in naming the date etc. The average of the 10 possible comparisons between 6 and 7 year boys shows that the latter averaged 16.1% higher than the former. The average increases in percentage passed from year to year are as follows,—6 to 7=16.1%; 7 to 8=13.5% 8 to 9=8.7%; 9 to 10=11.2%; 10 to 11=6.0%; and 11 to 12=0.2%. These figures show strikingly the irregularity of the growth from age to age. Comparing these average percentage increases in tests with the averages shown in Tables 9 and 11, there is no observable relation between this increase and the increase in average age from age to age, or the increase in average grade from age to age. The smallest increase in the tests (0.2%, 11 to 12) coincides with the smallest increase in average age from year to year (0.84 yr.), and the smallest increase in average grade from year to year. The other relations are varied.

The fact of the variability in the results of the non-selected boys stands out. The irregularity of the growth of the various abilities, and the fact that in 21% of the cases the boys of one age are actually lower than those of the previous age, point to the conclusion that certain allowances will have to be made for chance variations. It is not possible to acccount for the variations in growth by reference to the relative increase in average age or average grade from year to year.

The results of the 165 non-selected girls, shown in italics in the first seven columns of table 13, were studied in the same manner as the results of the boys. 60 differences between the percentage performance of the girls of each age and those of the preceding age were obtained. These differences ranged from —33% to +50%, the median being 7% (Q=8%). 10 of the deviations (17%), were minus values. The largest deviations were shown in the tests of naming 60 words, (—33%, 11 to 12), counting stamps (—20%, 9 to 10), and drawing designs

(—14%, 8 to 9). The remaining 7 minus deviations were below 10%.

The average increases in the percentage passed from year to year are as follows,— 6 to 7=3.9%; 7 to 8=15%; 8 to 9=8.8%; 9 to 10=10.1%; 10 to 11=8.7%; 11 to 12=1.8%. Both boys and girls show the smallest average increase in the percentage passed in the step from 11 to 12, and the magnitudes of the increases agree fairly well except for the step from 6 to 7. The increase of the 7 year girls over the 6 year girls is 3.9%, the next to the smallest increase of one age group over any preceding group. The 7 year boys, however, show an average increase of 16.1%, over the 6 year boys, the largest increase of any group of boys over any preceding group. It will be difficult, then, to draw conclusions concerning sex differences from a comparison of the 6 year boys and girls, for the 6 year girls are either a superior selection or the 6 year boys are an inferior selection if the character of these groups be judged by the comparison with the 7 year subjects. The same comparison, on the other hand, might indicate that the 7 year girls were an inferior selection and the 7 year boys a superior selection from the general run. It is only possible to point out the irregularity, however, it is not possible to show the cause of the irregularity.

A comparison of the average increase in the percentage passed by girls from age to age with the increase in the average ages shown in Table 9 shows no demonstrable relation to exist. Comparing this growth in the ability on the tests with the growth in average grade, shown in Table 11, shows a very positive relation to exist between these factors. Where the increase in average grade is smallest (i.e. from 6 to 7 and from 11 to 12), the increase in the tests is smallest (3.9% and 1.8%), while the greatest increase in grade (from 9 to 10 and from 7 to 8) coincide with the greatest increase in the test abilities (10.1% and 15.0%). This relation was not indicated in the results of the boys. The explanation of this fact that a correlation between the increase in the tests with grade was found in the results of the girls but not of the boys is a matter of speculation. It has been shown that the boys have a higher variability in grade than

girls. This tendency of the boys to be distributed in a wider range of grades might nullify the grade correlation slightly, but probably not to any considerable extent. The fact that the causes of this variation are not determined serves to illustrate the dangers of comparing the results of two groups when the factors operating on the groups are not known.

The foregoing study of the growth of the various abilities from age to age in each sex, and the analysis of the causes influencing this growth, demonstrates the great variability of the results. This fact of variability must be considered before drawing conclusions concerning sex differences by the method of comparing the results of boys and girls of each age.

The percentage differences between the performance of non-selected boys and girls of each age are shown in Table 13. In actual magnitude, these differences vary from 0% to 36%, the median being 9% (Q=5.5%). 75% of the differences are 17% or under, and only 16% are over 20%. In regard to sign, the differences vary from —36% to +26%, the median being —3.5% (Q=8.75%), showing a slight general superiority of the girls. If the number of possibilities of variation in comparing the results of small groups of non-selected subjects are taken into consideration, the presence of mental defectives, of subjects having language difficulties, of subjects in different grades influenced by different training, the possibility of a superior selection of subjects at one age group than at another, and the probability that similar chance samplings would not fall at the same age, the fact of correspondence indicated in Table 13 has more meaning than the fact of divergence.

The variability indicated in the study of the growth of abilities with age was so great that it makes interpretation of the results in terms of sex differences very difficult, and warranted conclusions impossible. It is legitimate to expect that the older subjects of either sex should make higher scores than the younger subjects of the same sex, but this was not found to be the universal rule. The boys' results showed minus deviations in 21% of the cases and the girls' results showed minus deviations in 17% of the cases. In one case the 12 year girls were 33% lower than

the 11 year girls. If this value (33%) be taken as the error due to chance variation, then only one value, that of —36%, (naming the months, age 12), may be taken as significant, and it has been seen that in this test the 12 year boys are 10% lower than the 9 year boys. The conclusion would follow, then, that there were no sex differences. This alternative, however, seems to place too much weight on one variation so that the truth probably lies in the assertion that the sex differences, that actually exist, are slight.

A study of the reactions of selected groups of boys and girls should throw light on the results from non-selected subjects, and make conclusions more certain. Subjects were selected by a process of elimination and selection. All of the subjects that were in the special class and minus grades were eliminated, along with all children of non-English speaking parents. From the following group of English speaking subjects in the regular grades all subjects were eliminated who had entered grade at an age very much above or below that of the general run of entrants.[4] The remaining subjects ranged in age from 4.3 years to 14.4 years, but were found to group rather closely around certain ages. It was possible to find four groups of boys and girls of approximately the same chronological ages. The character of these subjects is indicated in Table 14.

The four groups of subjects, chronologically from 6.0 to 6.9, 7.6 to 8.9, 9.7 to 10.9 and 11.7 to 13.3 (which will be referred to as 6, 8, 10 and 12), were distributed in approximately the same grades, and had approximately the same average age and average grade. The results of these groups are shown in Table 15, which is arranged to show all the facts for selected subjects that were given for non-selected subjects in Tables 12 and 13. The first four columns show the percentage that each test was given to each group. The next four columns show the percentage or the proportion that the subjects in each group passed each

[4] The ages on entering each grade of the subjects retained were as follows,—Kindergarten = 4, 5 and 6; Grade I = 5, 6 and 7; Grade II = 6, 7 and 8; Grade III = 8, 9 and 10; Grade IV = 9, 10 and 11; Grade V = 10, 11 and 12; Grade VI = 11, 12 and 13.

TABLE 14

Age in Grade Distribution, Average Grade and Average Age of 167 Selected Subjects. 86 Boys and 81 Girls.

		Age in Grade Distribution									
Age Group	Sex	K	I	II	III	IV	V	VI	TOTAL	Average Grade (M.V.)	Average Age (M.V.)
6.0 to 6.9	Boys	5	13						18	0.72 (0.40)	6.52 (0.22)
	Girls	3	13	2					18	0.89 (0.39)	6.53 (0.22)
7.6 to 8.9	Boys		7	13	3				23	1.83 (0.51)	8.09 (0.38)
	Girls		2	13	5				20	2.15 (0.43)	8.32 (0.38)
9.7 to 10.9	Boys				6	12	2		20	3.80 (0.48)	10.37 (0.36)
	Girls				9	7	5		21	3.81 (0.69)	10.14 (0.32)
11.7 to 13.3	Boys					2	8	15	25	5.52 (0.58)	12.35 (0.55)
	Girls					3	8	11	22	5.36 (0.64)	12.41 (0.46)

test. Column A shows the total number of times each test was given to all boys and girls, Column B, the weighted average age (the average ages given in Table 14 being used), and Column C the percentage that all subjects passed each test. The next four columns show the percentage that the boys are above (+) or below (—) the girls. Column D (derived from Column B), gives the difference between the average ages of all subjects to whom each test was given. Column E (derived from Column C), gives the differences between the percentages passed by all boys and girls on each test.

The growth of the various abilities with age in the selected groups of subjects is more uniform than that shown by the non-selected subjects. Only three cases appear in which the younger subjects make higher scores than those of older subjects, these exceptions occurring in the tests of describing pictures (—3%, girls 6 to 8), naming colors (—7%, girls 6 to 8), and naming months (—9%, boys, 10 to 12). In the comparison of the sexes 41 differences are obtained varying in magnitude from —28% to +26%, the median being 0% (Q=9.5%). In actual magnitude the differences vary from 0 to 28, the median being 10% (Q=4.75%), the median being 1% higher than that of non-selected data, and the variability 0.75% less. 75% of the differences were less than 14%.

TABLE 15

Results of 167 Selected Subjects. (86 Boys and 81 Girls).

		Percentage test was given.				Percentage or proportion passed				Columns			Percentage that boys are higher or lower than girls				Columns	
		6	8	10	12	6	8	10	12	A	B	C	6	8	10	12	D	E
Counting 13 pennies.	Boys	100	87	15	0	94	95	3/3		41	7.57	95	— 6	— 5			—.25	— 5
	Girls	*100*	*70*	*24*	*5*	*100*	*100*	*5/5*	*1/1*	*38*	*7.82*	*100*						
Describing pictures.	Boys	100	83	20	0	100	100	4/4		41	7.38	100	+11	+14			—.44	+10
	Girls	*100*	*70*	*24*	*5*	*89*	*86*	*5/5*	*1/1*	*38*	*7.82*	*90*						
Copying a diamond.	Boys	100	83	40	0	56	100	8/8		45	7.87	82	—27	+ 8			—.16	— 8
	Girls	*100*	*60*	*43*	*5*	*83*	*92*	*9/9*	*1/1*	*40*	*8.03*	*90*						
Naming four colors.	Boys	100	83	15	0	89	95	3/3		40	7.55	93	—11	+ 2			—.27	— 4
	Girls	*100*	*70*	*24*	*5*	*100*	*93*	*5/5*	*1/1*	*38*	*7.82*	*97*						
Counting from 20 to 0.	Boys	94	96	75	12	6	36	93	3/3	57	8.45	46	—27	—28	+ 2		+.10	—16
	Girls	*83*	*70*	*52*	*9*	*33*	*64*	*91*	*2/2*	*42*	*8.35*	*62*						
Counting stamps.	Boys	78	100	75	12	14	48	93	3/3	55	8.50	55	+ 1	+ 8	+26		—.08	+11
	Girls	*89*	*75*	*57*	*23*	*13*	*40*	*67*	*5/5*	*48*	*8.58*	*44*						
Repeating all digits.	Boys	100	100	100	100	16	39	56	75	86	9.53	48	—11	0	—11	+ 4	+.01	— 4
	Girls	*100*	*95*	*100*	*100*	*27*	*39*	*67*	*71*	*80*	*9.52*	*52*						
Naming days of week.	Boys	100	100	75	12	44	78	100	3/3	59	8.41	75	—23	+ 9	0		+.14	— 2
	Girls	*100*	*80*	*57*	*9*	*67*	*69*	*100*	*2/2*	*48*	*8.27*	*77*						
Giving day and date.	Boys	67	96	90	60	8	27	100	100	67	9.37	60	—12	—12	0	0	+.03	— 5
	Girls	*83*	*90*	*86*	*68*	*20*	*39*	*100*	*100*	*66*	*9.34*	*65*						
Naming the months.	Boys	56	91	85	52	10/1	29	94	85	61	9.38	56		—12	+ 6	—15	—.20	—12
	Girls	*39*	*85*	*76*	*55*	*7/2*	*41*	*88*	*100*	*52*	*9.58*	*67*						
Naming pieces of money.	Boys	39	74	100	72	7/2	24	60	78	62	9.88	52		— 7	+ 8	— 4	+.03	+ 2
	Girls	*50*	*65*	*100*	*77*	*9/1*	*31*	*52*	*82*	*60*	*9.85*	*50*						
Copying designs from memory.	Boys	11	61	100	68	2/0	14/3	30	53	53	10.26	34			+ 6	—10	+.08	— 1
	Girls	*28*	*45*	*100*	*73*	*5/0*	*9/3*	*24*	*63*	*51*	*10.18*	*35*						
3 words in sent. two ideas.	Boys	6	57	100	100	1/1	13/6	75	88	59	10.64	75			— 4	—12	+.40	— 2
	Girls	*33*	*50*	*100*	*91*	*6/0*	*10/9*	*71*	*100*	*57*	*10.24*	*77*						
3 words in sent. one idea.	Boys	6	57	100	100	1/1	13/2	50	68	59	10.64	51			+17	+23	+.40	+18
	Girls	*33*	*50*	*100*	*91*	*6/0*	*10/3*	*33*	*45*	*57*	*10.24*	*33*						
Naming 60 words in 3 minutes.	Boys	11	30	80	96	2/0	7/3	69	88	49	10.85	71			+13	+20	+.52	+18
	Girls	*39*	*25*	*76*	*86*	*7/0*	*5/3*	*56*	*68*	*47*	*10.33*	*53*						
Giving rhymes with 3 words.	Boys	6	22	65	96	1/0	5/3	69	71	43	11.12	67			+ 6	— 3	+.49	+ 1
	Girls	*17*	*30*	*76*	*86*	*3/1*	*6/4*	*63*	*74*	*44*	*10.63*	*66*						
Defining by use.	Boys	44	65	55	16	8/8	15/15	11/11	4/4	38	8.87	100					+.29	+ 6
	Girls	*61*	*55*	*43*	*14*	*11/10*	*11/10*	*9/9*	*3/3*	*34*	*8.58*	*94*						
Defining superior to use.	Boys	44	65	55	16	8/1	15/9	11/4	4/4	38	8.87	47						
	Girls	*61*	*55*	*43*	*14*	*11/2*	*11/2*	*9/5*	*3/0*	*34*	*8.58*	*26*					+.29	+21

The change of the median of the series of differences from —3.5% (non-selected) to 0% (selected) shows that the elimination of over age and special grade pupils has helped the boys more than the girls, and has altered the general relations between the sexes. This fact is also indicated by the average difference in the percentages that all subjects pass each test, the average for non-selected subjects being —1.4% and for selected subjects +1.6%. The non-selected boys from 6 to 12 were given, in all, 2436 tests, these tests being passed 60.8% of the time. The non-selected girls were given 2195 tests, passing 61.6%, the advantage being 0.8% in their favor. The selected boys were given 1125 tests, passing 64.3%, an advantage of 0.1% over the girls who passed 64.2% of 1034 tests. The foregoing changes indicate clearly that the selection of subjects has changed the general relations between the sexes, helping the boys more than the girls.

The relations between the results of selected and non-selected subjects may be studied by a comparison of the differences between the percentages passed by all subjects. If the differences between the scores of the boys and girls are due to but one factor, that of sex differences, then the correlation between the two methods of study should be very nearly absolute. The correlation (Pearson product-moments formula) between the differences in the percentage passed by all boys and girls according to the two methods is 0.726 (pe=0.075). This correlation between the two methods is high, but it would probably be high inasmuch as the 167 selected subjects are included in the 352 non-selected subjects. The results of the two methods show certain large discrepancies. The changes of the greatest magnitude are those shown by the 60 words test (+4% by the first method to + 18% by the second), the tests of defining in terms superior to use (+7% to +21%), of naming the days of the week, (—16% to —2%), giving rhymes, (—10% to +1%), naming colors, (—14% to —4%), copying the diamond, (+1% to —8%), and counting from 20 to 0 (—8% to —16%). The comparison of the median differences shows that the selected method tends to improve the results of the boys more

than those of the girls. All of the changes in the results of the two methods are not in favor of the boys, however, the total scores on the diamond and 20 to 0 tests showing changes in favor of the girls. If the cause of the variations shown by the first method is the presence of a few children of non-English speaking parents, to special class and minus grade children, then the elimination of this source of error should change the results in only one direction.

The analysis of the results of selected subjects, therefore, does not lessen the difficulty of the interpretation of the results in the light of sex differences. The rate of growth of the various abilities with age is irregular. The analysis of the irregularities points to the fact that the boys or girls of any age may be a chance selection of superior or inferior subjects at that age. The method of comparing selected subjects would tend to eliminate the inferior selection of subjects, but would not eliminate the possibility of a superior selection.

The comparison of the results of the sexes shows differences at certain ages and on certain tests that are as high as 20%. The problem involved is that of deciding whether these large differences are due to chance or to differences in the reactions of the sexes. Certain tests show large deviations first in favor of one sex and then in favor of the other. If a difference of a percentage of any magnitude on any test is to be attributed to a sex difference, then the same line of reasoning will show that in certain tests the abilities change from one sex to the other. The analysis of the tests that show this crossing of ability should throw light on the other tests.

Three tests show substantial differences in favor of both sexes according to both methods. In the test of copying the diamond, the non-selected girls lead at the start, age 6, and the boys are ahead at 7, 8 and 9, the same relations being shown by selected subjects of 6 and 8. In the test of copying the designs from memory, the non-selected girls are 24% below the boys at age 9 and 21% above the boys at age 12, the same relations being shown by the selected subjects of 10 and 12. In the test of naming 60 words in three minutes, the non-selected girls are

19% above the boys at 9, and 19% below at 12. The selected boys of 10 and 12 are in advance of the girls in this test.

These three tests are crucial in the consideration of the problem of whether differences shown between the boys and girls are due to actual sex differences or due to accidental causes. Each of these tests may be studied by a method more accurate than that of comparing the percentage passed at each age. The reproductions of the diamond were arbitrarily sorted in six groups according to their merits by a method described in the discussion of the personal equation. The first group contained the best reproductions, the sixth, the poorest. The reproductions of the designs were graded from 0 to 20 by an arbitrary point system described under the discussion of the personal equation. A measure of the ability in the 60 word test is the actual number of words given in three minutes, a measure recorded by the experimenters in each case. Table 16 shows the average score made by the non-selected and selected boys and girls of each age in these three tests.

TABLE 16

Average Score (Mean Variation) of Subjects of Each Age on Three Tests.

		Copying the Diamond Average Group of the Reproductions.		Drawing the Designs Average number of points scored.		Naming 60 words Average number of words given in three minutes.	
		Boys	Girls	Boys	Girls	Boys	Girls
unselected subjects	6	4.27(1.28)	3.57(1.24)				
	7	2.85(1.04)	3.17(1.37)				
	8	2.20(1.15)	3.24(1.57)	8.06(6.19)	9.00(5.25)		
	9	2.33(0.89)	3.00(1.29)	10.29(5.30)	5.32(4.61)	52.93(11.20)	59.91(10.10)
	10			9.17(5.33)	9.18(6.73)	68.12(13.12)	61.76(11.25)
	11			8.64(6.73)	10.94(7.06)	73.65(13.35)	71.28(14.25)
	12			8.64(6.02)	11.08(6.08)	68.75(12.28)	58.14(12.57)
selected subjects	6	4.27(1.20)	3.33(1.26)				
	8	2.32(1.00)	3.00(1.17)				
	10			9.55(5.60)	7.29(6.42)	67.31(12.74)	62.13(11.39)
	12			12.53(5.38)	13.56(5.55)	75.33(10.92)	66.84(13.87)

The relations indicated by the percentage passed are also indicated by the more reliable method of comparing the average scores. In the test of copying the diamond, the 6 year non-selected girls average 0.70 group better than the boys, while the

selected girls are 0.94 ahead. The comparison of the 7, 8 and 9 year subjects shows the boys ahead in all cases, the 8 year non-selected boys averaging over one group higher. The non-selected boys show an improvement of two groups from 6 to 9, while the girls show an improvement of only half a group. One sex shows a decided growth of ability, the other practically none. If the differences indicated are to be taken as real, it will be necessary to assume that the girls pick up the ability to draw a diamond easier than the boys, but that this ability once obtained remains constant—that the effect of maturity operates on one sex but not on another. The number of cases on which this assumption is based (174 subjects from 6 to 9) is so small, and the chances of variation in the selection of subjects of different intellectual status in each age group is so large, that the assumption is not substantiated.

The relations indicated in the test of copying the designs are more variable than those of the diamond test. The 9 year non-selected boys show an improvement over the 8 year boys, but from 9 to 12 there is a gradual decrease in the ability, so that the 11 and 12 year boys are only slightly ahead of the 8 year boys. The relations shown by the non-selected girls are exactly the reverse of those of the boys. The 9 year girls are very much lower than the 8 year girls, and a gradual increase appears from 9 to 12 instead of a decrease. The comparison of these opposite relations gives a maximum difference in favor of the boys at 9 and the girls at 12. If the relations indicated in this test are to be considered definite, the assumption is involved that the influence of increasing age on one sex is exactly opposite to that on the other sex, an assumption that is not substantiated in view of the small number of cases (183 subjects from 8 to 12) and the possibility of selecting subjects of chance superiority in the small groups at each age.

The relations indicated in the test of naming 60 words are more constant than those shown in the diamond or design tests. Both sexes show a growth of ability from 9 to 11 and a decrease from 11 to 12. The growth is irregular, however, the girls showing less growth from 9 to 10, and a greater drop from 11 to

12, so that a comparison of the sexes shows a deviation in favor of the girls at 9 and of the boys at 12. The assumption of any large sex differences in this test involves the assumption that 12 year girls have less ability in this test than 9 year girls, and that the influence of maturity operates differently on the two sexes, an assumption that is not substantiated in view of the many variable factors.

The conclusion that a definite crossing of ability between the sexes occurs in the tests of copying the diamond, copying designs and naming 60 words, is not substantiated. It is not justifiable to attribute a difference of 20% between the sexes to a real sex difference on one test and not on another. If the differences shown between the results of the sexes in the tests of constructing a sentence containing one idea, of naming the months, naming the days of the week, counting stamps and naming colors are to be attributed to sex differences, then the variations in ability shown in the diamond, design and 60 word test must be assumed to be definite. These assumptions were not found to be substantiated, however, so that it is not possible to draw any conclusions concerning sex differences from a study of the percentage that selected or unselected subjects of each age pass each test.

The variable influences due to the selection of subjects of different status at each age are eliminated or counterbalanced to some extent by combining the subjects of all ages. The differences between the percentages that all boys and girls pass each test are to some extent influenced by the ages of the subjects to whom each test was given. The correlation (Pearson product-moments formula) of the differences between the percentages that all non-selected boys and girls passed each test with the difference between the average ages of all the non-selected boys and girls to whom each test was given is 0.394 (pe=0.134). The correlation between the same arrays from selected subjects (i.e. between Columns D and E of Table 15) is 0.388 (pe=0.135). These correlations between the tests and age are high enough to indicate that the factor of age is present to some extent. The close correspondence in the correlations from the two methods

indicates that the age factor is present to the same extent in both methods. The tests vary in the degree with which they correlate with age, so that it is not possible to estimate the amount of the influence of this factor. Furthermore, it has been seen that the results from the two methods are not in strict accordance, that the elimination of inferior subjects caused changes in the results in both directions. For these reasons, it is not possible to draw any conclusions concerning sex differences from a comparison of the percentages passed by all subjects.

Certain negative conclusions are, however, possible. The number of subjects at each age in both methods is comparatively small. The chances of variations due to factors other than sex differences has been shown to be very large. The fact of correspondence between the results of the two sexes is therefore of more importance than the fact of divergence. 75% of the differences between the non-selected boys and girls are 17% or under, while the same proportion of the differences between selected boys and girls falls under 14%. If it is assumed that the subjects of any age should not test lower than those of any preceding age, and allowance is made for differences between the sexes that are exaggerated on account of the chance falling off of ability with older subjects, only 9% of the differences between the non-selected boys and girls are over 20% (derived from Table 13).

The evidence from the foregoing methods of study points to the conclusion that the sex differences, if present, are under 20% or 25% as a maximum, and that deviations of this magnitude are marked exceptions to the general run of differences. The conclusion that the differences that might possibly be attributed to the sex factor are slight, has no meaning unless the word "slight" is defined independently of the writer's personal opinion. The differences shown between the results of the sexes are smaller than those that were attributed to the factor of the personal equation in the study of the results of the four experimenters. It was concluded that certain tests were influenced by grade training. These tests showed from 40% to 60% improvement from one grade to another, so that the greatest influence that may be attributed to the sex factor is only approximately

one half that due to grade training. The following study of the diagnostic value of the tests will show that the deviations that might be attributed to the sex factor are insignificant when compared to the differences between the reactions of normal and retarded children to the individual tests.

Most of the investigators who have studied the factor of sex differences in the Binet tests, have studied them from the standpoint of the "mental ages" or total scores made by the subjects of both sexes. A few investigators have studied sex differences in the light of the individual tests. Descoeudres (20) reports the results of the application of the Binet tests to 24 subjects, one good and one poor pupil of each sex from each of six school grades, drawing conclusions from this investigation concerning the diagnostic value of the individual tests and the sex differences involved. Obviously the number of subjects is too small to allow any conclusions to be drawn. Chotzen (18) compared the percentage that all feeble-minded boys and girls passed each of 15 tests, finding differences varying in magnitude from 1% to 20%. The largest deviations were those of 20% in favor of the boys in the test of copying the diamond, 13% in favor of the girls in the test of executing three commissions, 12% in favor of the boys in naming the pieces of money, 11% in favor of the girls in the test of repeating a sentence of 16 syllables, and 10% in favor of the girls in detecting omissions in pictures. All other differences were less than 10%.

Bloch and Preiss (9) examined 155 normal Volkschule children (79 boys and 76 girls) varying in age from 7 to 13. Bobertag's translation was used. These investigators found very striking differences in the reaction of the sexes to the individual tests, the differences running as high as 52%, most of them in favor of the boys. The differences between the performances of the boys and girls of each age were calculated, without reference to the many sources of variation. The factor of the personal equation is not treated, and this factor alone might cause these variations. If a more careful analysis of the results had been made, it is very probable that the conclusions would have been modified to some extent. The fact that the 11 year

boys are 37% higher than the 11 year girls on the test of criticising absurdities is most certainly modified by the fact that the 11 year subjects are 30% lower than the 10 year subjects in the test of repeating 7 digits. The small number of subjects (in five cases less than 10), would tend to emphasize chance variations. The fact that the number of subjects is too small to warrant definite conclusions is pointed out by the authors. Stern (62) in commenting upon these results, points out the significance of the fact that the inferiority of the girls extends to so many different kinds of tests. The results of Bloch and Preiss are in almost complete contradiction to the results of the present investigation. They find large differences, and find practically all of these differences in favor of the boys. This investigation shows a general run of differences very much smaller, and a slight general superiority of the non-selected girls. The mere fact of contradiction in the results of the two investigations would indicate that the differences were not produced by the common factor of sex. Rogers and McIntyre (54) give no figures, but report that they have studied their results in the light of sex differences, and have found no correlation between their results and those of Bloch and Preiss.

The results of the investigators who have compared the "mental ages" or total scores of children of different sexes are somewhat at variance. Goddard (30) reports that there are more backward boys than girls. Stern notes that Goddard's results do not bear out his statement, for the percentage of boys and girls testing two or more years retarded is the same (18.5%). The accuracy of Goddard's statement depends on the criterion[5] used for measuring backwardness. Although Goddard's state-

[5] If the criterion is four or more years retarded, there are more backward boys than girls (boys = 3.7%, girls = 3.1%). If the criterion is three or more years backward, there are more girls than boys (boys = 8%, girls = 9.1%). If the criterion is two or more years backward, the proportions are the same, as Stern notes. If the criterion is one year or more retarded, there are more backward boys than girls (boys = 41.4%, girls = 35.6%). There are more girls than boys testing at and above age according to Goddard's results. 34.7% of the boys and 36.6% of the girls test at age, while 23.8% of the boys and 27.7% of the girls test one year or more above age.

ment concerning the backwardness of the boys may be interpreted differently, his figures leave no doubt concerning the fact that there are more girls than boys at and above age, and therefore indicate a general superiority of the girls.

Bobertag (10) computed the average "mental age" of 90 boys and 90 girls regularly distributed from 7 to 12. The subjects were selected according to school grades, so that the average grade of each group differed by exactly one grade. His results show the boys ahead 0.06 yr. at 7, 0.14 yr. at 8 and 9, 0.20 yr. at 10, 0.19 yr. at 11 and 0.14 yr. at 12. These findings cannot be considered entirely out of harmony with those of Goddard, for, as this investigation shows, there may be a change in the relation of non-selected boys and girls and selected boys and girls.

Yerkes and his co-workers (82), scoring some of the Binet tests according to the point system, show that the girls of English speaking parents are superior to the boys of the same parentage between 5 and 7, that they fall below with minor variations till 11, where they again surpass the boys at 12 and 13, falling below at 14 and 15. The differences between the sexes are smaller and of less practical importance than the differences due to the language factor, but the authors suspect "that at certain ages serious injustice will be done to individuals by evaluating their scores in the light of norms which do not take account of sex differences." (page 73). In contradiction to these results are those of Terman and his co-workers (67), who, scoring the Stanford revision of the Binet scale according to "intelligence quotients," find differences of but 2% to 4% in these quotients in favor of the girls, and who conclude from the basis of their studies of sex differences that the conclusions of Yerkes are unjustified. These two investigations used tests different in character and differently weighted, so that the results would not necessarily have to correspond.

The one common feature of most of the researches on sex differences in the Binet-Simon tests is that the differences are small. Burt and Moore (17) summarize the work of various investigators in the general field of sex differences, and report an investigation of their own on 67 boys and 63 girls, 12½ to 13½

years of age. They discuss their results and those of the other authors in the order of the complexity of the mental processes involved. They find a high correlation between the size of the sex difference and the simplicity of the capacities compared—the higher the process, and the more complex the capacity, the smaller the sex difference.

The general trend of the investigations on sex differences indicates that no very large differences are to be expected in the application of intelligence tests, and that the differences to be expected will vary according to the nature of the tests. The results of this investigation are in agreement with the general trend of the investigations in showing only slight differences that might be attributed to the sex factor. The results do not show on what tests, if any, these differences occur. Conclusions concerning the amount of influence of this factor must be drawn from more exhaustive investigations on the individual tests. The research of Bateman (3), for instance, is conclusive in the test of naming colors. Bateman shows that there is a difference of 14% in favor of the girls in this test, showing furthermore that the factor of school training causes an improvement of but 18%. The results would indicate that the test should be placed in the fifth or sixth year, but the sex difference of 14% would probably not warrant the placing of the test in a different age group for boys and girls.

The investigations of Bolton (11) and Wooley (79) would show that small differences in favor of the girls are to be expected in the tests of repeating digits, and possibly in all memory tests. The investigations of Gilbert (27), Thompson (68), Burt and Moore, and Peterson and Doll (51) would indicate that a slight difference in favor of the boys should appear in the test of arranging five weights. Ruger's (55) finding of striking differences in favor of men in a series of puzzle tests, and Wooley and Fisher's finding of large differences in favor of the boys in the Healy puzzle-box test would show that rather large differences might appear in the general class of "puzzle" tests.

Even though the sex differences in intelligence tests may be shown to be small, scientific procedure should demand that the

investigator who standardizes any test or system of tests should treat his results in such a way as to demonstrate that the factor is present or not present. The burden of proof should still be on the person who maintains that sex differences are not involved. The knowledge of sex differences is especially important in diagnosing border-line cases of mental defect, where the diagnosis must often be made on the qualitatively different character of the responses to individual tests.

VI. SUMMARY.

One of the fundamental assumptions in the construction of the Binet-Simon scale is the correlation of the individual tests with age. The correlation of the tests with age is affected by the error due to incomplete data, by the influence of the personal equation of the experimenter, and by the training the subject has received in school.

The influence of the personal equation of the experimenter was found to be more marked in some tests than in others, the influence being most marked in the tests of copying the diamond, indicating omissions in pictures, defining in terms superior to use, drawing designs from memory, detecting absurdities in statements and reconstructing dissected sentences.

The variations between the experimenters could be traced to three sources,—

1) to the use of apparatus, variations in which were due to,
   a) the construction of the test material, and
   b) the use of alternative questions;

2) to the technique of the experimenters in giving the tests; and

3) to observation errors made by the experimenters in marking a response passed or failed.

It is possible to eliminate all three sources of error.

The effect of school training was more marked on some tests than on others, the effect being most marked in the tests of counting stamps, counting backward from 20 to 0, enumerating the days of the week and the months, giving the day and the date, naming the pieces of money, making change, and reconstructing dissected sentences. Tests that involve school training should be standardized on a different basis than those relatively independent of this factor.

Although the comparison of "mental ages" and pedagogical ages gives no information concerning the general correlation be-

tween the Binet tests and the school grades, the study of the individual tests establishes the fact of a general correlation.

The correlation of the individual tests with grade is higher than the correlation of the tests with age, this fact being indirect evidence of the value of the tests as measures of intelligence.

Sex differences were found to be slight as compared with the influence due to the personal equation or grade training.

Since variations occur in the results due to the influence of the personal equation and grade training, certain allowances must be made for these factors in making diagnoses on the basis of the tests. The scale is therefore a qualitative rather than a quantitative instrument.

The investigator who wishes to use his results for standardizing age norms should use only those data based on the complete method of experimenting, and should treat his results in such a way as to demonstrate the presence or absence of the variable factors of the personal equation, grade training and sex differences.

BIBLIOGRAPHY

1. ABELSON, A. R. The Measurement of Mental Ability of "Backward" Children. Brit. J. of Psychol., 1911, 4, 268-314.
2. AYRES, L. P. The Binet-Simon Measuring Scale of Intelligence: Some Criticisms and Suggestions. Psychol. Clinic, 1911, 5, 187-196.
3. BATEMAN, W. G. The Naming of Colors by Children. Ped. Sem., 1915, 22, 469-486.
4. BINET, A. Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'ecole. Année psychol., 1911, 17, 145-201.
5. BINET, A. AND SIMON T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. Année psychol., 1905, 11, 191-244.
6. BINET, A. AND SIMON T. Application des méthodes nouvelles au diagnostic du niveau intellectual chez des enfants normaux et anormaux d'hospice et d'ecole primaire. Année psychol., 1905, 11, 245-336.
7. BINET, A. AND SIMON T. Le développement de l'intelligence chez les enfants. Année psychol., 1908, 14, 1-94.
8. BINET, A. AND SIMON T. La mesure du développement de l'intelligence chez les jeunes enfants. Bull. de la soc. libre pour l'etude psychol. de l'enfant. 1911, 11, 187-256.
9. BLOCH, E. AND PREISS, A. Ueber intelligenzprüfungen an normalen Volksschulkindern nach Bobertag. (Methode von Binet und Simon) Zsch. f. agnew. Psychol., 1912, 6, 539-547.
10. BOBERTAG, O. Ueber Intelligenzprüfungen (nach der Methode von Binet und Simon). I. Methodik und Ergebnisse der einzelnen Tests. Zsch. f. angew. Psychol., 1911, 5, 105-203. II. Gesamtergebnisse der Methode. Zsch. f. angew. Psychol., 1912, 6, 495-537.
11. BOLTON, T. L. The Growth of Memory in School Children. Amer. J. of Psychol., 1892, 4, 362-380.
12. BONSER, F. G. The Reasoning Ability of Children of the Fourth, Fifth and Sixth School Grades. New York: Columbia Univ., 1910, pp. 133.

13. BRIDGMAN, O. Mental Deficiency and Delinquency. J. of Amer. Med. Assoc., 1913, 61, 471-472.
14. BRIGHAM, C. C. An Experimental Critique of the Binet-Simon Scale. J. of Educ. Psychol., 1914, 5, 439-448.
15. Buffalo conference. J. C. Bell, C. S. Berry, W. S. Cornell, E. A. Doll, J. E. W. Wallin, G. M. Whipple, Informal Conference on the Binet-Simon Scale: Some Suggestions and Recommendations. J. of Educ. Psychol., 1914, 5, 95-100.
16. BURT, C. Experimental Tests of General Intelligence. Brit. J. of Psychol., 1910, 3, 94-177.
17. BURT, C. AND MOORE, R. G. The Mental Differences between the Sexes. J. of Exp. Ped., 1912, 1, 273-284, 355-388.
18. CHOTZEN, F. Die Intelligenzprüfungsmethode von Binet-Simon bei schwachsinnigen Kindern. Zsch. f. angew. Psychol., 1912, 6, 411-494.
19. DECROLY, O. AND DEGAND J. La mesure de l'intelligence chez des enfants normaux d'après les tests de M. Binet et Simon: nouvelle contribution critique. Arch. de psychol., 1910, 9, 81-108.
20. DESCOEUDRES, A. Les tests de Binet et Simon et leur valeur scolaire. Arch. de psychol., 1911, 11, 331-350.
21. DESCOEUDRES, A. Exploration de quelques tests d'intelligence chez des enfants anormaux et arriérés. Année psychol., 1911, 11, 351-375.
22. DOLL, E. A. Inexpert Binet Examiners and their Limitations. J. of Educ. Psychol., 1913, 4, 607-609.
23. DOUGHERTY, M. L. Report on the Binet-Simon Tests given to 483 Children in the Public Schools of Kansas City, Kansas. J. of Educ. Psychol., 1913, 4, 338-352.
24. DRESSLAR, F. B. Studies in .the Psychology of Touch. Amer. J. of Psychol., 1894, 6, 313-368.
25. EBBINGHAUS, H. Ueber eine neue Methode zur Prüfung geistigen Fahigkeiten und ihre Andwendung bei Schulkindern. Zsch. f. Psychol. 1897, 13, 401-459.
26. FERNALD, W. E. The Diagnosis of the Higher Grades of Mental Defect. Amer. J. of Insan., 1914, 70, 741-752.
27. GILBERT, J. A. Researches on the Mental and Physical Development of School Children. Stud. fr. Yale Psychol. Lab., 1894, 2, 40-100.
28. GODDARD, H. H. The Binet-Simon Measuring Scale for Intelligence. (Revised edition) Vineland, N. J. The Training School, 1911, pp. 16.

29. GODDARD, H. H. Standard Method of giving the Binet Test. Training School, 1913, 10, 23-32.
30. GODDARD, H. H. Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence. Ped. Sem., 1911, 18, 232-259.
31. GODDARD, H. H. Three Annual Testings of 400 Feeble-Minded Children and 500 Normal Children. Psychol. Bull. 1913, 10, 75-77.
32. HAINES, T. H. Diagnostic Value of some Performance Tests. Psychol. Rev., 1915, 22, 299-305.
33. HEALY, W. The Individual Delinquent. Boston: Little Brown & Co., pp. 830.
34. HEALY, W. AND FERNALD, G. M. Tests for Practical Mental Classification. Psychol. Monog. 1911, 13 (No. 54) pp. 53.
35. HUEY, E. B. The Binet Scale for Measuring Intelligence and Retardation. J. of Educ. Psychol., 1910, 1, 435-444.
36. HUEY, E. B. A Point Scale of Tests for Intelligence. Baltimore: Warwick & York (folder) 4 pp.
37. KATZENELLENBOGEN, E. W. A Critical Essay on Mental Tests in their Relation to Epilepsy. Epilepsia, 1913, 4, 130-173.
38. KITE, E. S. The Binet-Simon Measuring Scale of Intelligence. Philadelphia: Committee on Provision for the Feeble-Minded, Bull. no. 1, pp. 29.
39. KITE, E. S. The Development of Intelligence in Children. (Contains translations of nos. 5, 6, and 7). Vineland, N. J.: The Training School (Publications of the Department of Research, No. 11), 1916, pp. 328
40. KITE, E. S. The Intelligence of the Feeble-Minded. (Translation of three articles by Binet and Simon on Feeble-mindedness) Vineland, N. J.: The Training School, (Publications of the Department of Research, No. 12), 1916, pp. 328.
41. KOHS, S. C. The Binet-Simon Measuring Scale of Intelligence: an Annotated Bibliography. J. of Educ. Psychol., 1914, 5, 215-224, 279-290. 335-346.
42. KOHS, S. C. The Practicability of the Binet Scale and the Question of the Borderline Case. Training School, 1916, 12, 211-224.
43. KUHLMAN, F. Some Results of Examining a Thousand Public School Children with a Revision of the Binet-Simon Tests of Intelligence by Untrained Examiners. J. of Psycho-Asthenics, 1914, 18, 233-269.

44. MARTIN, A. L. A Contribution to the Standardization of the De Sanctis Tests. Training School, 1916, 13, 93-110.
45. MEUMANN, E. Vorlesungen zur Einführung in die experimentelle Pädagogik und ihre psychologischen Grundlagen. Leipzig: W. Englemann 1913, Vol. II, pp. 800.
46. MEUMANN, E. Ueber eine neue Methode der Intelligenzprüfung und über den Wert der Kombinationsmethoden. Zsch. f. päd. Psychol. und exp. Päd., 1912, 13, 145-163.
47. MORROW, L. AND BRIDGMAN, O. Delinquent Girls Tested by the Binet Scale. Training School, 1912, 9, 33-36.
48. NORSWORTHY, N. The Psychology of Mentally Deficient Children. New York: (Columbia Univ. thesis) 1906, pp. 111.
49. OTIS, A. S. Some Logical Aspects of the Binet Scale. Psychol. Rev. 1916, 23, 129-152, 165-179.
50. OTIS, M. The Binet Tests Applied to Delinquent Girls. Psychol. Clinic, 1913, 7, 127-134.
51. PETERSON, A. M. AND DOLL, E. A. Sensory Discrimination in Normal and Feeble-Minded Children. Training School, 1914, 11, 110-118, 135-144.
52. PILLSBURY, W. B. The Psychology of Reasoning. New York: D. Appleton & Co., 1910, pp. 304.
53. PYLE, W. H. A Psychological Study of Bright and Dull Pupils. J. of Educ. Psychol., 1915, 6, 151-156.
54. ROGERS, A. L. AND MCINTYRE, J. L. The Measurement of Intelligence in Children by the Binet-Simon Scale. Brit. J. of Psychol., 1915, 7, 265-299.
55. RUGER, H. A. Sex Differences in the Solution of Mechanical Puzzles. (In report of New York branch of American Psychological Assoc.) J. of Phil., Psychol., etc., 1914, 11, 412-413.
56. SCHMITT, C. The Binet-Simon Tests of Mental Ability. Ped. Sem. 1912, 19, 186-200.
57. SCHMITT, C. Standardization of Tests for Defective Children. Psychol. Monog., 1915, 19 (No. 83) pp. 181.
58. SIMPSON, B. R. Correlations of Mental Ability. New York: Columbia Univ., 1912, pp. 122.
59. SMITH, F. O. The Effect of Training in Pitch Discrimination. Univ. Iowa Stud. in Psychol., Vol. VI. Psychol., Monog., 1914, 16 (No. 69) 67-103.
60. STENQUIST, J. L., THORNDIKE, E. L. AND TRABUE, M. R. The Intellectual Status of Children who are Public Charges. Arch. of Psychol. 1915. 33, pp. 52.

61. STERN, W. Die differentielle Psychologie in ihren methodischen Grundlagen. Leipzig: Barth, 1911, pp. 503.
62. STERN, W. The Psychological Methods of Testing Intelligence. (Whipple, G. M., trans. fr. German) Educ. Psychol. Monog., No. 13, Baltimore: Warwick & York, 1914, pp. 160.
63. Symposium on Mental Tests. (Conducted by C. E. Seashore under "Communications and Discussions") J. of Educ. Psychol., 1916, 7. (R. M. Yerkes, 163-164).
64. TERMAN, L. M. Genius and Stupidity. Ped. Sem., 1906, 13, 307-373.
65. TERMAN, L. M. The Measurement of Intelligence. Boston: Houghton Mifflin Co., 1916, pp. 362.
66. TERMAN, L. M. AND CHILDS, H. G. A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence. J. of Educ. Psychol., 1912, 3, 61-74, 133- 143, 198-208, 277-289.
67. TERMAN, L. M., LYMAN, G., ORDAHL, G., ORDAHL, L., GALBREATH, N. AND TALBOT, W. The Stanford Revision of the Binet-Simon Scale, and some Results from its Application to One Thousand Non-Selected Children. J. of Educ. Psychol., 1915, 6, 551-562.
68. THOMPSON, H. B. Psychological Norms in Men and Women. Chicago: Univ. of Chicago Press, 1903, pp. 188.
69. THORNDIKE, E. L. The Significance of the Binet Mental Ages. Psychol. Clinic, 1914, 8, 185-189.
70. THORNDIKE, E. L. An Introduction to the Theory of Mental and Social Measurements. New York: Teachers' College, 1913, pp. 277.
71. THORNDIKE, E. L., LAY W. AND DEAN, P. R. The Relation of Accuracy in Sensory Discrimination to General Intelligence. Amer. J. of Psychol., 1909, 20, 364-369.
72. TOWN, C. H. A Method of Measuring the Development of The Intelligence of Young Children. (Authorized translation of no. 8) Lincoln, Ill.; Courier-Herald Co. 1913, pp. 82.
73. WALLIN, J. E. W. Experimental Studies of Mental Defectives. Educ. Psychol. Monog. No. 7. Baltimore, Warwick & York, 1912, pp. 155.
74. WITMER, L. On the Relation of Intelligence to Efficiency. Psychol. Clinic, 1915, 9, 61-86.
75. WHIPPLE, G. M. Manual of Mental and Physical Tests. Baltimore: Warwick & York, 1910, pp. 534.

76. WHIPPLE, G. M. Manual of Mental and Physical Tests. Baltimore: Warwick & York, 1914, pp. 690, 2 vol.
77. W[HIPPLE], G. M. The Amateur and the Binet-Simon Tests. J. of Educ. Psychol., 1912, 3, 118-119.
78. W[HIPPLE], G. M. Amateruism in Binet Testing once more. J. of Educ. Psychol., 1913, 4, 301-302.
79. WOOLEY, H. T. A New Scale of Mental and Physical Measurements for Adolescents and some of its Uses. J. of Educ. Psychol. 1915, 6, 521-550.
80. WOOLEY, H. T. AND FISHER, C. R. Mental and Physical Measurements of Working Children. Psychol. Monog. 1914, 18 (No. 77) pp. 247.
81. WYATT, S. The Quantitative Investigation of Higher Mental Processes. Brit J. of Psychol., 1914, 6, 109-133.
82. YERKES, R. M., BRIDGES, J. W. AND HARDWICK, R. S. A Point Scale of Measuring Mental Ability. Baltimore: Warwick & York, 1915, pp. 213.

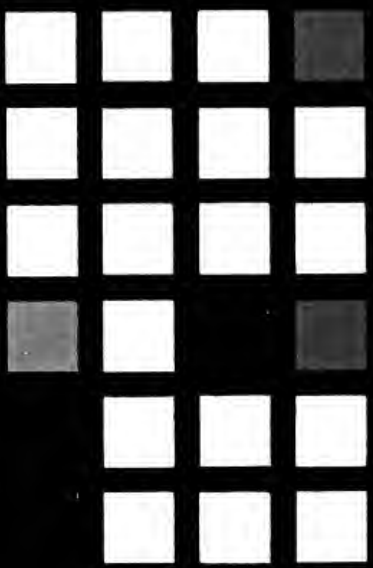
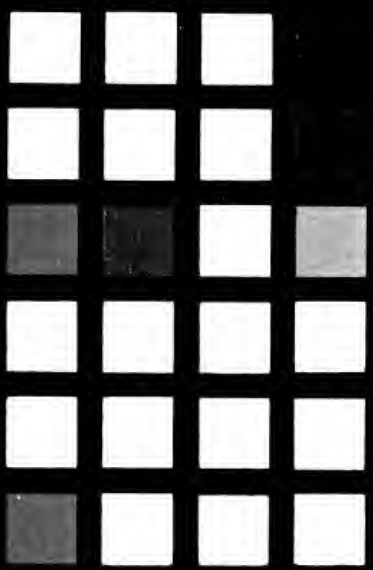
THIS BOOK IS DUE ON THE LAST DATE STAMPED BELOW

AN INITIAL FINE OF 25 CENTS

WILL BE ASSESSED FOR FAILURE TO RETURN THIS BOOK ON THE DATE DUE. THE PENALTY WILL INCREASE TO 50 CENTS ON THE FOURTH DAY AND TO $1.00 ON THE SEVENTH DAY OVERDUE.

JUL 29 1943

LD 21–100*m*-7,'39 (402s)

YD 22588

UNIVERSITY OF CALIFORNIA LIBRARY